

**DAS URTEIL VON ALLGEMEINPRAKTIKERN
ALS KRITERIUM DER GÜLTIGKEIT VON
TESTFRAGEN FÜR DIE ÄRZTLICHE PRÜFUNG**

Christian Schönhals

1974



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41551922_0067

Aus dem Institut für Ausbildungsforschung München
Leiter: Dr.med. Hannes Kapuste

DAS URTEIL VON ALLGEMEINPRAKTIKERN ALS KRITERIUM DER
GÜLTIGKEIT VON TESTFRAGEN FÜR DIE ÄRZTLICHE PRÜFUNG

Inaugural-Dissertation
zur Erlangung der Doktorwürde in der gesamten Medizin
verfaßt und einer Hohen Medizinischen Fakultät
der Ludwig Maximilians-Universität zu München

vorgelegt von
Christian Schönhals
aus
München

1974

Gedruckt mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter: Prof. Dr. med. K. Betke

Dekan: Prof. Dr. W. Spann

Tag der mündlichen Prüfung: 10. 9. 1974



Inhaltsverzeichnis

	Seite
1. Schriftliche Leistungsmessungen in der medizinischen Ausbildung	1
.1 Leistungstests als Prüfinstrumente	1
.2 Leistungstests in der Ausbildungsforschung und zur Kontrolle von Lernzuwachs	2
2. Die drei Hauptkriterien in der Testtheorie	4
.1 Das Kriterium der Objektivität	4
.2 Das Kriterium der Reliabilität	5
.21 Retest-Reliabilität	5
.22 Paralleltest-Reliabilität	6
.23 Testhalbierungsverfahren	7
.3 Das Kriterium der Validität	8
.31 Kriterienbezogene Validität	9
.32 Inhaltliche Gültigkeit	11
.321 Unterrichtsbezogene inhaltliche Gültigkeit	11
.322 Berufszielbezogene inhaltliche Gültigkeit	12
3. Der Tripod Test	13
.1 Entstehung des Tripod Test	13
.2 Objektivität des Tripod Test	14
.3 Reliabilität des Tripod Test	14
.4 Gültigkeit des Tripod Test	15
.41 Kriterienbezogene Gültigkeit	15
.42 Inhaltliche Gültigkeit	16
4. Problemstellungen zum Experiment	19
.1 Auswahl der Sachverständigen	19
.2 Durchführbarkeit der Befragung	20
.3 Zuverlässigkeit der Ergebnisse	21
.4 Welche neuen Ergebnisse sind zu erwarten	21
5. Methodik	22
.1 Befragungsmodus	22

	Seite
.2 Auswertung der erhaltenen Daten	23
.21 Berechnung der durchschnittlichen Bewertung der einzelnen Fragen und der gesamten Tests aus den Mittelwerten der Gutachterurteile	23
.22 Berechnung von Durchschnitt und Verteilung der Bewertungen bei den einzelnen Gutachten	24
.23 Berechnung der Korrelationen der gewonnenen Gutachterurteile mit den vorhandenen Professorenurteilen sowie der Urteile der Gutachter untereinander	24
.24 Überprüfung von Zusammenhängen zwischen der durchschnittlichen Bewertung einzelner Fragengruppen ("Subtests") und ihrer Zuordenbarkeit zu verschiedenen inhaltlichen und formalen Kategorien	24
6. Ergebnisse	27
.1 Rücklauf	27
.2 Ergebnisse der Auswertung	28
.21 Durchschnitt und Verteilung der Bewertungen bei einzelnen Gutachtern	28
.22 Durchschnittliche Bewertung der einzelnen Fragen und des gesamten Tests	31
.23 Korrelationen der Gutachterurteile mit Professorenurteilen sowie der Urteile der Gutachter untereinander	35
.24 Zusammenhänge zwischen der Bewertung von Fragen und ihre kategorielle Zuordnung zu verschiedenen Subtests	37
.241 Positiv bewertete Fragenkategorien	43
.242 Negativ bewertete Fragenkategorien	50
.243 Fragenkategorien, für die sich in den beiden Testformen entgegengesetzte Bewertungstendenzen abzeichnen	58
.244 Fragenkategorisierung nach Herkunftsfächern	65
.245 Fragenkategorisierung nach formalen Kriterien	68
7. Diskussion der Ergebnisse	72
.1 Beteiligung der Ärzte	72
.2 Beurteilung der Tests und der Fragen	72
.3 Zuverlässigkeit der Gutachterurteile	74

	Seite
.4 Zusammenhänge zwischen kategorieller Zuordnung und Gutachterurteilen	75
.41 Inhaltskategorien	75
.42 Formalkategorien	76
8. Zusammenfassung	79
9. Literaturverzeichnis	83

Das Urteil von Allgemeinpraktikern als Kriterium der Gültigkeit von Testfragen für die ärztliche Prüfung

1. Schriftliche Leistungsmessungen in der medizinischen Ausbildung

Die Verwendung schriftlicher Prüfungsverfahren hat sich in den vergangenen Jahren für Semesterabschlußprüfungen u.ä. immer mehr eingebürgert. Mit Inkrafttreten der neuen Approbationsordnung für Ärzte vom 28.10.70 wird nun auch der weitaus größte Teil der offiziellen medizinischen Prüfungen schriftlich abgehalten. Als Frageform ist das Multiple-Choice-Verfahren vorgeschrieben, bei dem der Prüfling anzugeben hat, welche der mit den Fragen vorgelegten Antworten er für zutreffend hält (17). Neben ihrer Funktion als Prüfinstrumente finden medizinische Leistungstest in der Ausbildungsforschung ihre Anwendung.

1.1 Leistungstest als Prüfinstrumente

Die Anwendung schriftlicher Leistungstests im Prüfungswesen bringt sicherlich verschiedene Vorteile gegenüber den herkömmlichen mündlichen Prüfungen. Ein sehr wesentlicher Gesichtspunkt ist dabei die Ökonomisierung des Prüfungsvorganges. Bei der vorgeschriebenen bundeseinheitlichen Durchführung der Prüfung wird es möglich sein, sämtliche Prüflinge eines Termins mit ein und demselben Test und den gleichen Fragen zu prüfen. Dadurch wird der erhebliche Aufwand, den die Herstellung schriftlicher Prüfungsfragen erfordert, auf das Mindestmaß beschränkt und betrifft zudem nurmehr diejenigen Hochschullehrer und anderen Spezialisten, die an der Erarbeitung der Prüfungsfragen in einem Zentralen Prüfungsfrageninstitut unmittelbar beteiligt sind. Für die übrigen Hochschullehrer stellt die Entlastung von den Prüfungsverpflichtungen die dringend notwendig gewordene Voraussetzung für die Intensivierung und Verbesserung des Unterrichtsangebotes dar.

Ein wesentlicher Faktor, der dabei jedoch überhaupt nicht erfaßt wird, ist die Persönlichkeit des Prüflings, die sicherlich für seine Eignung zum ärztlichen Beruf nicht belanglos ist. Inwieweit eine dreistündige praktisch-persönliche Prüfung von jeweils vier Kandidaten, wie sie für

denSchluß der Ausbildung nach der neuen Approbationsordnung vorgesehen ist, ausreichen kann, dieses Manko auszugleichen mag dahingestellt sein.

Sicherlich ist zumindest in dieser Hinsicht das Staatsexamen bisheriger Prägung die überlegene Prüfung gewesen. Auch von den Studenten selbst wird offenbar die mündliche bzw. kombiniert mündlich-schriftliche Prüfung ganz überwiegend bevorzugt. Bei Interviews mit Studenten sprachen sich von den Befragten nur drei für rein schriftliche, sechs für rein mündliche und 30 für eine Kombination von mündlichen und schriftlichen Prüfungen aus (4).

Wenn aber, wie dies durch das Inkrafttreten der neuen Approbationsordnung der Fall ist, rein schriftliche Prüfungen den allergrößten Teil der ärztlichen Prüfungen ausmachen, so ist zu fordern, daß bei der Konstruktion der Tests für diese Prüfungen qualitative Kriterien optimal erfüllt werden, damit wenigstens der kognitive Anteil ärztlicher Kompetenz möglichst zuverlässig gemessen werden kann. Auf diese zu erfüllenden qualitativen Kriterien wird weiter unten noch näher einzugehen sein.

1.2 Leistungstests in der Ausbildungsforschung und zur Kontrolle von Lernzuwachs

Neben der Verwendung als selektive Prüfungen können schriftliche Leistungstests noch eine Reihe anderer Funktionen im Rahmen der medizinischen Ausbildung und deren Reform erfüllen. So können beispielsweise Leistungskontrollen durch Tests, die ohne selektive Funktion in den Ausbildungsgang regelmäßig einbezogen werden, allen Beteiligten wichtige Informationen liefern. Die Anwendung der EDV bei der Auswertung von Testergebnissen ermöglicht eine Beleuchtung der Ergebnisse unter verschiedenen Aspekten. So kann der Dozent als Grundinformation den Lehrerfolg aus der durchschnittlichen Leistung seiner Studenten entnehmen. Darüberhinaus kann er jedoch z.B. auch feststellen, ob bestimmte Ausbildungsinhalte besser oder schlechter verstanden wurden als der Durchschnitt und seinen Unterricht ggf. dementsprechend verändern. Die Studenten können ihrerseits neben der Grundinformation über ihre Gesamtpunktzahl Auskunft über ihre relative Leistung im Vergleich zum Klassendurchschnitt erhalten. Außerdem besteht ebenfalls die Möglichkeit, die Leistung jedes Studenten

nach Untergruppen der Ausbildungsinhalte aufzuschlüsseln. Auf diese Weise kann der Student erfahren, auf welchen Teilgebieten er möglicherweise hinter dem Durchschnitt seiner Kollegen zurückliegt und seine Arbeitsanstrengungen dann dort konzentriert vermehren. Diese Form der Unterrichts- und Lernkontrolle stellt die Voraussetzung dar, um trotz wachsender Studentenzahlen eine optimale Kooperation zwischen Lehrenden und Lernenden zu ermöglichen.

Des weiteren läßt sich mit Hilfe medizinischer Leistungstests die Effizienz verschiedener Unterrichtsmethoden miteinander vergleichen, indem man Testungen an verschiedenen Hochschulen mit unterschiedlichen Unterrichtsformen durchführt und die durchschnittlichen Leistungsniveaus untersucht und mit den Methoden in Beziehung setzt. Noch besser läßt sich der Erfolg bestimmter Unterrichtsangebote feststellen, indem man die Leistung der Studenten vor Beginn der Untersuchung mißt und nach Ablauf des Unterrichts den Leistungszuwachs durch erneute Testung bestimmt. So ließe sich beispielsweise unter einer Anzahl alternativer Unterrichtsmethoden diejenige mit dem größten Effizienzgrad ermitteln. In begrenztem Umfang mag auch bereits eine bundeseinheitliche Durchführung von Prüfungen Grundlagen für Vergleich und Reformen aufzeigen, die zu einer einheitlichen Verbesserung des Leistungsniveaus führen können, wie dies im Begleittext zur neuen Approbationsordnung erwartet wird (1).

Ein wichtiger Beitrag auf dem Wege der Verbesserung und Überprüfung der Ausbildungsmethoden und -ziele ist durch die Festlegung des Prüfungsstoffes in den Anlagen zur Approbation bereits geleistet worden. Allerdings werden einzelne Abschnitte und Formulierungen dieser Stoffkataloge noch ausreichend Anlaß zur Diskussion der einen oder anderen Auslegungsweise bieten. Solche Detaildiskussionen mit dem Ziel einer Eingrenzung und verbindlichen Definition der Ausbildungsziele sind heute bei der gebotenen Straffung und Beschränkung der Ausbildungsinhalte auf das wirklich Wesentliche wichtiger denn je. Im übrigen sind sie die unerläßliche Voraussetzung für die Herstellung gültiger, qualitativ zufriedenstellender Prüfungen.

Welches im einzelnen die Hauptkriterien sind, an denen die Qualität solcher Prüfungen zu messen ist, soll der folgende Abschnitt kurz darstellen.

2. Die drei Hauptkriterien in der Testtheorie

Auf medizinische Leistungstests sind, wie auf alle psychologischen Tests, zu denen diese gehören, bei der Qualitätsanalyse vor allem drei Kriterien anzuwenden, die als die drei Hauptkriterien in der Testtheorie bekannt sind: Objektivität, Reliabilität (Zuverlässigkeit) und Validität (Gültigkeit).

2.1 Das Kriterium der Objektivität

Unter der Objektivität eines Tests versteht man den Grad der Unabhängigkeit der Ergebnisse vom Untersucher. Einflüsse des Untersuchers oder der Untersucher - falls es sich um mehrere handelt - sind in verschiedenen Stadien des Testvorgangs denkbar, so daß man zwischen Durchführungs-, Auswertungs- und Interpretationsobjektivität unterscheidet. Bei medizinischen Leistungstests in Form von Antwort-Wahl-Fragen sind alle drei Objektivitätsarten praktisch 100% gewährleistet(13).

Auf der Stufe der Durchführung nämlich sind durch schriftliche Arbeitsanweisungen und Aufgabenstellungen Einflüsse durch die Person des Untersuchers auf die Testergebnisse ausgeschlossen. Die Auswertung der Testergebnisse erfolgt - zumindest bei Testungen in größerem Rahmen - auf elektronischem Wege über Markierungsbögen, von denen Computer die Lösungen unmittelbar ablesen können. Die Interpretation schließlich besteht in der Addition der richtigen Lösungen zum Testrohwert jedes Prüflings und wird ebenfalls vom Computer besorgt. Damit sind auch Auswertung und Interpretation allen Verfälschungen durch Einflüsse der Untersucher praktisch entzogen. Die Problematik der Objektivität taucht auf einer anderen Ebene auf.

Die Objektivität eines schriftlichen Leistungstests, bei dem Durchführung, Auswertung und Interpretation optimal objektiviert sind, wird im wesentlichen bestimmt durch die Objektivität der einzelnen Aufgabe. Diese wiederum ist eine Funktion von sprachlicher Form und inhaltlicher Richtigkeit. Für den speziellen Fall einer Mehrfach-Wahl-Aufgabe ist von einer objektiven Frage zu verlangen, daß sie das Problem, das sie zur Lösung stellt, beschreibt, ohne daß Mißverständnisse möglich sind, und daß eindeutig zu entscheiden ist, welche der zur

Auswahl stehenden Alternativen die richtige Antwort ist. In vielen Fällen können die Autoren eines Tests die Frage selbst soweit redigieren, daß ihre Objektivität evident ist. Bestehen Zweifel, vor allem bezüglich der sachlichen Richtigkeit oder Eindeutigkeit, so gilt als Kriterium für die Objektivität, daß sich mehrere Sachverständige über die richtige Beantwortung der Frage einig sein müssen (10).

2.2 Das Kriterium der Reliabilität

Ein weiteres wichtiges Kriterium für die Qualität eines Tests ist eine ausreichende Zuverlässigkeit oder Reliabilität. Man versteht darunter nach traditioneller Definition den Grad der Genauigkeit mit welcher der Test ein bestimmtes Merkmal mißt, unabhängig davon, ob er dieses Merkmal zu messen beansprucht (13). Dies bedeutet, daß die Kenntnis von Reliabilitätskennwerten allein ohne eine Untersuchung der inhaltlichen Relevanz eines Tests noch keine Rückschlüsse auf seine Qualität gestattet, weil sie nur zum Ausdruck bringen, wie genau man irgendwelche Merkmale der untersuchten Personen damit feststellen kann, nicht aber, welcher Art diese Merkmale sind; ob es sich also um diejenigen handelt, die der Test zu messen beansprucht.

Für die quantitative Bestimmung eines Reliabilitätskoeffizienten stehen unterschiedliche Methoden zur Verfügung. Welche davon im einzelnen verwandt wird, hängt von der Art des Testmaterials ab. Die drei klassischen Methoden der Reliabilitätskontrolle sind das Retestverfahren, die Paralleltestmethode und das Testhalbierungsverfahren. Eine Verbesserung und zugleich Vereinfachung des letzteren stellt heute die Anwendung der Formeln von Kuder-Richardson dar, mit deren Hilfe es möglich ist, ausgehend von den üblichen bei der Testanalyse gewonnenen Parametern (Anzahl Fragen, Durchschnitt und Standardabweichung des Tests und der Fragen) die Reliabilität des Tests mathematisch zu berechnen (7).

2.2.1 Retest-Reliabilität

Bei diesem Verfahren wird eine Stichprobe von Probanden in angemessenem Zeitabstand zweimal mit dem gleichen Test untersucht und anschließend die Übereinstimmung der beiden Messungen miteinander

als Reliabilitätskoeffizient bestimmt. Dies entspricht im Prinzip der in den Naturwissenschaften gebräuchlichen Methode, die Genauigkeit einer Messung durch Wiederholung zu bestimmen. Allerdings ist die Übertragung dieses Konzepts auf psychologische Meßmethoden nur in beschränktem Maße zulässig, weil bei diesen häufig eine Veränderung des Meßobjekts durch die Messung stattfindet. So z.B. auch bei Leistungstests, wo ganze Lösungen oder Lösungskonzepte aus der 1. Messung bei der erneuten Testvorlage aus dem Gedächtnis gefunden werden können. Man kann dem zu begegnen versuchen, indem man einen möglichst langen Zeitraum zwischen den beiden Testläufen verstreichen läßt. Dabei muß man aber zunehmend mit einer Veränderung des Informationsstandes der einzelnen Probanden rechnen. Da eine voll befriedigende Lösung dieses Problems kaum zu finden ist, hat die wiederholte Testdarbietung als Reliabilitätskontrolle für medizinische Leistungstests praktisch kaum eine Bedeutung.

Demgegenüber eignet sich jedoch der Versuchsplan der Testwiederholung, um mit Hilfe eines Tests, der hinsichtlich seiner Reliabilität und der übrigen Qualitätskriterien bereits überprüft ist, zu untersuchen, welchen Einfluß die erste Testung und insbesondere die Information über ihr Ergebnis auf das Lernverhalten der untersuchten Studenten hat.

Dabei ist v.a. auch daran zu denken, daß Studenten - durch schlechte Ergebnisse in bestimmten Teilbereichen des Tests motiviert - kompensatorische Lernanstrengungen in diesen Bereichen vollbringen und daraufhin bei der zweiten Testbearbeitung signifikant besser abschneiden. Es ist wohl kaum zu bezweifeln, daß es sich dabei um eine wünschenswerte Kenntnisvermehrung bei den Studenten handelt, während es unter Anlegung klassischer Reliabilitätskriterien eine mangelhafte Reliabilität des Tests ausweisen würde (9).

2.2.2 Paralleltest-Reliabilität

Dieses Verfahren setzt wie der Name bereits sagt, das Vorhandensein von zwei oder mehreren Parallelformen eines Tests voraus. Paralleltests müssen bestimmten teststatistischen Kriterien der Vergleichbarkeit oder Äquivalenz genügen, die u.a. eine Übereinstimmung in Inhalt und Schwierigkeit der Test erfordern (13). Zur Ermittlung der Reliabi-

lität wird dann eine Stichprobe von Probanden in zwei Testläufen beiden Tests unterzogen (bei mehr als zwei Tests entsprechend mehrere) und aus der Korrelation der Ergebnisse der Koeffizient für die Paralleltest-Reliabilität errechnet.

Durch die Verwendung von Paralleltests geht man der wesentlichen Fehlerquelle aus dem Wege, die bei der Retest-Reliabilitätskontrolle im Erinnern schon einmal gefundener Lösungen bestand. Durch eine Variation des Meßinstruments im Detail vermeidet man die Auswirkung der Veränderungen, die bei psychologischen Tests am Meßobjekt - der Merkmalsausprägung beim Probanden also - durch den Testvorgang hervorgerufen werden können. So wird diese Form der Reliabilitätskontrolle auch für medizinische Leistungstests anwendbar. Sie ist jedoch wegen des Aufwandes, den die Herstellung paralleler Testformen nach strengen statistischen Kriterien darstellt, im allgemeinen nur dann sinnvoll, wenn parallele Tests über das Verfahren der Reliabilitätsbestimmung hinaus benötigt werden, um sie beispielsweise in der Ausbildungsforschung verwenden zu können.

2.2.3 Testhalbierungsverfahren

Bei dieser Methode wird nun im Gegensatz zu den beiden vorher erwähnten nur eine Testung vorgenommen. Um trotzdem zwei Ergebnisreihen als Grundlagen einer Korrelationsberechnung zu erhalten, wird der Test nach der Bearbeitung in zwei Hälften gespalten und die Punktzahlen der Probanden in der einen Hälfte werden mit denen in der anderen korreliert. Auch hierbei muß eine möglichst hohe Vergleichbarkeit der beiden Testhälften angestrebt werden, was je nach Eigenart des Tests unterschiedliches Vorgehen bei der Testhalbierung erfordert. Besteht der Test aus verschiedenartigen Fragen unterschiedlicher Thematik, wie das bei medizinischen Leistungstests allgemein der Fall ist, so muß eine gleichmäßige Verteilung aller Fragentypen auf beide Testhälften erreicht werden. Als Ergebnis entstehen dann gewissermaßen zwei "Paralleltests" von halber Länge des Ausgangstests, für die sich in üblicher Weise ein Reliabilitäts-Koeffizient aus den Testergebnissen errechnen läßt.

Bei dieser Art der Reliabilitätskontrolle muß allerdings noch der Einfluß der Testlänge auf die Reliabilität beachtet werden. Die Reliabi-

tät wächst nämlich mit zunehmender Testlänge - vorausgesetzt natürlich, daß die hinzugefügten Fragen nicht von minderer Qualität sind. Deshalb wird sich der Reliabilitätskoeffizient für die beiden Testhälften niedriger berechnen als es dem Test in seiner gesamten Länge zukäme, für deren Anwendung er eigentlich konzipiert ist. Zur Korrektur dieser experimentell bedingten Unterschätzung der Testreliabilität stehen jedoch Formeln zur Verfügung, die es gestatten, dem Reliabilitätsgewinn bei Verdoppelung der Testlänge abzuschätzen (7).

Die Verwendung der Testhalbierungsmethode zur Reliabilitätsbestimmung bietet zwei wesentliche Vorteile. Sie erspart zum einen den Aufwand der Konstruktion von Paralleltests. Wenn auch die Testhalbierung - gerade auch bei medizinischen Leistungstests - sehr sorgfältig vorgenommen werden muß, so braucht doch die Prüfungsfragensammlung, die das notwendige Ausgangsmaterial für die Testkonstruktion darstellt, praktisch nur halb so umfangreich zu sein. Zum zweiten wurde bereits erwähnt, daß die Testhalbierungsmethode nur eine einmalige Testdurchführung erfordert, was ebenfalls eine Verringerung des Arbeits- und Zeitaufwandes zur Folge hat.

Der Vorgang der Testhalbierung selbst ist jedoch kompliziert und mit Fehlerquellen behaftet, weshalb die eleganteste Methode der Reliabilitätsbestimmung in der Anwendung der von Kuder-Richardson entwickelten Formel 20 auf der Basis der üblicherweise bekannten Testparameter besteht (7). Sie vermeidet die inhaltliche Schwierigkeit einer äquivalenten Testhalbierung, indem sie den Durchschnitt aller denkbaren Testhalbierungen in die Berechnung des Reliabilitätskoeffizienten einbezieht (9).

2.3 Das Kriterium der Validität

Während Objektivität und Reliabilität als eher formale Kriterien der Testbeurteilung aufgefaßt werden können, betrifft das Kriterium der Gültigkeit oder Validität die inhaltliche Qualität eines Tests. Ein Test gilt dann als valide, wenn er tatsächlich das mißt, was er messen soll. Ob dies der Fall ist, muß vor der Anwendung eines Tests durch eine Validitätsprüfung festgestellt werden. Zwei wichtige Konzepte der Validität sind die kriterienbezogene und die inhaltliche Validität (13).

2.3.1 Kriterienbezogene Validität

Psychologische Tests dienen dazu, den Ausprägungsgrad bestimmter Merkmale bei den untersuchten Personen zu messen und soweit wie dies möglich ist, in quantitativen Resultaten festzuhalten. Dies gilt auch für medizinische Leistungstests, bei denen man das Merkmal, dessen Ausprägungsgrad quantitativ erfaßt werden soll, mit dem Oberbegriff "ärztliche Kompetenz" bezeichnen könnte. Um überprüfen zu können, ob ein solcher Test tatsächlich ärztliche Kompetenz mißt, kann man andere Informationen über die Ausprägung des gleichen Merkmals, ein "Außenkriterium", gewinnen. Dieses "Außenkriterium" kann verschiedener Art sein. Es kann sich dabei um einen anderen bereits erprobten medizinischen Leistungstest handeln oder um Noten aus einer mündlich-persönlichen Prüfung, welche die ärztliche Kompetenz der Kandidaten bewerten. Die Validitätskontrolle verläuft nun dergestalt, daß man eine Stichprobe von Versuchspersonen sowohl dem zu prüfenden Test unterzieht, als auch mit dem Außenkriterium bewertet. Das Maß der Übereinstimmung der Bewertungen durch den Test mit denen durch das Außenkriterium, das mittels einer Korrelationsberechnung errechnet wird, läßt sich in einem dimensionslosen numerischen Wert angeben, der als Validitätskoeffizient bezeichnet wird und gewöhnlich im Bereich zwischen 0 und + 1 liegt.

Die Errechnung eines numerischen Validitätskoeffizienten ist wünschenswert, da sie beispielsweise den Vergleich verschiedener Testentwürfe oder auch die Festlegung von Mindestanforderungen erleichtert. Andererseits ist die Validitätskontrolle nach dem Konzept der kriterienbezogenen Gültigkeit mit relativ großen Schwierigkeiten der Durchführung verbunden. Deshalb beschränkt man sich in der Praxis häufig mit dem weiter unten zu besprechenden Konzept der inhaltlichen Gültigkeit, das u. a. jedoch nicht unmittelbar zu vergleichbaren numerischen Koeffizienten führt.

Das Problem bei der Anwendung des Konzepts der kriterienbezogenen Validität liegt in der Validität des Außenkriteriums selbst bzw. der Erstellung eines validen Außenkriteriums. Ebenso wie an der Validität des Tests muß auch an der Gültigkeit des zweiten Tests oder der mündlichen Prüfung gezweifelt werden, solange sie nicht überzeugend nachgewiesen werden kann. In diesem Fall müßte konsequenterweise

ein neues Außenkriterium erstellt werden, um anhand dessen die Validität des ersten Außenkriteriums zu überprüfen. Auf diese Weise müßte man dann beliebig lang fortfahren. Es ist jedoch zulässig, diesem unendlichen Regress auszuweichen, indem man die Validität des Außenkriteriums durch kompetente Sachverständige in einem sog. "Rating"-Verfahren abschätzen läßt und die Validität dieser Urteile als voraussetzbar betrachtet. Muß man letztlich ohnehin den wissenschaftlich exakt reproduzierbaren Bereich verlassen, so kann man nötigenfalls auch von vornherein auf ein Außenkriterium verzichten und die Validität des Tests direkt durch ein Sachverständigen-Rating ermitteln.

Die Erstellung eines Außenkriteriums ist allerdings dann von eigenem Interesse, wenn sie zur Überprüfung der prognostischen Valenz eines Tests erfolgt. Viele Eignungs- oder Leistungstests dienen nicht nur dazu, den aktuellen Ausprägungsgrad eines Merkmals zu messen, sondern es wird für sie auch eine prognostische Valenz oder Voraussagegültigkeit beansprucht. Dies gilt auch für die ärztlichen Prüfungen, von denen man Voraussagegültigkeit für die zu erwartende Qualität ärztlicher Verrichtungen bei den Examenskandidaten erwartet. Um zu überprüfen, ob diese Art der Gültigkeit tatsächlich vorhanden ist, müßte man ein Außenkriterium erstellen, das den Ausprägungsgrad des prognostizierten Merkmals, also beispielsweise der praktisch-ärztlichen Qualität bei einer früher mit dem Test untersuchten Gruppe von Ärzten gültig repräsentiert. Die Gültigkeit eines solchen Außenkriteriums, das in irgendeiner Form die Qualität ärztlicher Leistungen zu messen imstande sein müßte, könnte man durch Sachverständige beurteilen lassen, um den erwähnten unendlichen Regress der Validierungen zu umgehen. Ein Validierungsprozeß, der eine Untersuchung der Korrelation zwischen den Leistungen in einem berufsqualifizierenden Test und der tatsächlichen Qualität späterer Berufspraxis erlauben würde, müßte von großem wissenschaftlichen und praktischen Interesse sein. Er wäre allerdings mit einem nicht unerheblichen Planungs- und Arbeitsaufwand verbunden, abgesehen von den Schwierigkeiten, die mit einer objektiven Messung praktisch-ärztlicher Qualität verbunden sind.

2.3.2 Inhaltliche Gültigkeit

Der größere technische Aufwand, der mit der kriterienbezogenen Validierung verbunden ist, führt - wie bereits erwähnt - in der Praxis häufig zur Beschränkung auf die inhaltliche Gültigkeitskontrolle. Ein Test gilt dann als inhaltlich valide, wenn seine Aufgaben oder Fragen eine repräsentative Stichprobe der zu prüfenden Merkmale darstellen (10). Eine ärztliche Prüfung ist hiernach valide, wenn sie die zu prüfenden Bereiche ärztlicher Kompetenz sowohl vollständig als auch in einem ihrer unterschiedlichen Bedeutung gerecht werdenden Verhältnis berücksichtigt. Ob dies der Fall ist, entscheidet bei einer Validierung nach dem Konzept der inhaltlichen Gültigkeit im allgemeinen das Urteil einer Gruppe kompetenter Sachverständiger.

Es ist offensichtlich, daß dieses Validierungsverfahren einfacher in der Durchführung ist als die kriterienbezogene Validitätskontrolle. An Nachteilen steht demgegenüber die größere Subjektivität und schlechtere quantitative Vergleichbarkeit der Sachverständigenurteile. Wie vergleichbar die Urteile Sachverständiger sind, hängt mit davon ab, wie übereinstimmend die Auffassungen von einer "repräsentativen Stichprobe der zu prüfenden Merkmale" sind, an welcher der Test gemessen werden soll. Dies wiederum setzt voraus, daß man sich einigt, unter welchem Gesichtspunkt man die repräsentative Stichprobe definiert: Unter dem einer unterrichtsbezogenen oder dem einer berufszielbezogenen inhaltlichen Gültigkeit.

3.2.1 Unterrichtsbezogene inhaltliche Gültigkeit

Bezieht man den Inhalt einer Prüfung oder eines Tests ausschließlich auf tatsächlich stattgefundenen Unterricht, so wird eine Übereinstimmung der Auffassungen über die für seine Gültigkeit anzulegenden Maßstäbe nicht allzu schwer zu erzielen sein. In diesem Sinne gültig ist ein Test dann, wenn er alle Gegenstände, die in den der Prüfung vorausgegangenen Unterrichtsveranstaltungen enthalten waren, berücksichtigt und zwar jeweils entsprechend der Ausführlichkeit und dem Raum, den sie tatsächlich darin eingenommen hatten. Hierfür lassen sich durch Unterrichtsprotokolle leicht objektive Anhaltspunkte geben, auf die sich alle Sachverständigen stützen können.

Selbstverständlich können die Ergebnisse einer unterrichtsbezogenen gültigen Prüfung immer nur so wertvoll sein, als der ihnen zu Grunde liegende Unterricht. Für eine berufsqualifizierende Prüfung wie die ärztliche Prüfung wäre die Anwendung dieses Konzepts also nur dann gerechtfertigt, wenn ihr ein Unterrichtsprogramm zugrunde läge, dessen erfolgreiche Absolvierung eine optimale Berufsvorbereitung bereits garantiert. Dies kann jedoch gerade in einem Stadium der Umbildung des Ausbildungsprozesses im Zuge einer Ausbildungsreform kaum vorausgesetzt werden. Deshalb sollte für die Validierung von Tests für die ärztliche Prüfung dem Konzept der berufszielbezogenen Gültigkeit der Vorzug gegeben werden.

2.3.2.2 Berufszielbezogene inhaltliche Gültigkeit

Mißt man die Gültigkeit einer Prüfung daran, wie repräsentativ sie für ein Berufsziel ist, das man zum Maßstab wählt, so bezieht man bis zu einem gewissen Grade auch die zuvor erwähnte Voraussagegültigkeit, die von dem Test erwartet wird - sofern es sich um einen berufsqualifizierenden oder Eignungstest handelt - in seine Beurteilung mit ein.

Die Gültigkeit der Prüfung oder des Tests wird daran bemessen, für wie geeignet man die in ihm enthaltenen Aufgaben hält, zu erwartende berufliche Qualität bei den Prüflingen im voraus abschätzen zu lassen. Dies setzt voraus, daß eine Übereinstimmung bei den Sachverständigen darüber besteht, wie das Berufsziel, das ihnen als Beurteilungsmaßstab dient, zu definieren ist: für welche Leistungen der Kandidat sich also qualifizieren muß und welche durch einen Test feststellbaren Kenntnisse notwendige und charakteristische Vorbedingungen hierfür sind. D.h. bei einer inhaltlich komplexen Prüfung wie der ärztlichen Prüfung auch, daß ein Konsens bei den Sachverständigen darüber bestehen muß, welchen Stellenwert die einzelnen Kenntnisbereiche innerhalb des gesamten Sachkataloges einnehmen, wie also eine repräsentative Stichprobe aller erforderlichen Kenntnisse beschaffen sein muß, um so der Definition eines inhaltlich gültigen Tests zu genügen.

In der vorliegenden Untersuchung soll ein medizinischer Leistungstest, der Tripod Test (11), durch sachverständige, praktische Ärzte begut-

achtet werden. Auf diese Weise soll geprüft werden, ob der Test eine berufszielbezogene inhaltliche Gültigkeit beanspruchen kann. Zunächst soll jedoch der verwendete Tripod Test näher beschrieben werden.

3. Der Tripod Test

Der Tripod Test, revidierte Fassung, ist ein standardisierter medizinischer Leistungstest, der im Mai 1969 vom Institut für Ausbildungsforschung München für wissenschaftliche Zwecke veröffentlicht wurde. Er besteht aus zwei Paralleltests (Form A und Form B) mit je 120 Fragen. Seine Entwicklung wurde von der Stiftung Volkswagenwerk finanziert (9).

3.1 Entstehung des Tripod Test

Die ursprüngliche Fassung des Tripod Test bestand in einer Auswahl von 441 Fragen aus deutschen, englischen und französischen Prüfungsfragensammlungen, welche die Leistung am Ende der medizinischen Ausbildung messen sollten. Die Auswahl erfolgte nach einer gewichteten Liste von Themenbereichen aus den Fächern Innere Medizin, Chirurgie, Kinderheilkunde, Gynäkologie und Geburtshilfe, Pathologie und Pharmakologie. Nach einem Zufallsverfahren wurden die Fragen auf zwei Tests mit 220 bzw. 221 Fragen verteilt. Da eine internationale Verwendung des Tests beabsichtigt war, wurden englische, französische und niederländische Übersetzungen hergestellt. Die Richtigkeit sowohl der Fragen als auch ihrer Übersetzungen wurden von Experten und mithilfe einer Reihe von statistischen Kriterien überprüft (9). Die so entstandenen Tests wurden u. a. in 21 Experimenten in sechs verschiedenen Ländern erprobt, an denen mehr als 1700 Medizinstudenten kurz vor ihren Abschlußexamina teilnahmen, 1330 Studenten bearbeiteten dabei eine der deutschen Fassungen, 410 eine (oder beide) englischen. Aufgrund der Analyse der bei den Experimenten gewonnenen Daten und Informationen wurden für die revidierte Form des Tests 81 Fragen ausgeschieden, sodaß der Tripod Test in seiner heutigen Form zu je 120 Fragen in beiden Testformen resultierte. Ein Teil der erhaltenen Fragen wurde bei der Revision korrigiert.

3.2 Objektivität des Tripod Test

Die Fragen des Tripod Test sind sämtlich Mehrfach-Wahl Aufgaben des einen oder anderen Typs (17). Damit erfüllen sie die formale Voraussetzung objektiver Befragung. Die Objektivität im Sinne einer Unmißverständlichkeit und Eindeutigkeit der richtigen Lösung wurde in Zusammenarbeit mit etwa 65 deutschen Fachspezialisten sowie anhand von einschlägigen Lehrbüchern überprüft und soweit nötig durch Korrekturen verbessert. Diese Art der Objektivitätskontrolle entspricht der eingangs angeführten Definition der Fragenobjektivität, die eine Übereinstimmung mehrerer Sachverständiger bezüglich der Richtigkeit einer Frage postuliert. Auch bei der Testdurchführung ist die Objektivität gewährleistet, da die Anleitungen zur Bearbeitung des Tests schriftlich sind und somit für alle Teilnehmer und zu jedem Zeitpunkt gleich. Die Auswertung der Testergebnisse geschieht maschinell und ist absolut objektiv. Der Kandidat markiert auf speziellen Antwortkarten jeweils welche Lösung er für richtig hält. Diese Markierungen können direkt vom Computer gelesen und ausgewertet werden. Das Testergebnis besteht in einer Punktzahl, die der Anzahl richtiger Markierungen entspricht.

3.3 Reliabilität des Tripod Test

Die folgende Tabelle zeigt die in verschiedenen Experimenten ermittelten Reliabilitätskoeffizienten berechnet nach der Formel 20 von Kuder-Richardson.

Tabelle 1: Die Reliabilitätskoeffizienten des Tripod Test berechnet nach Experimenten an verschiedenen deutschen und ausländischen Universitäten.

	HH	G	Tü	Ffm	Mü	Lüb	HD	Be	Chic	E _s	L	K	E _i
Form A	.91	.91	.92	.91	.72	.84	.94	.88	.84	.84	.90	.83	.86
Form B	.91	-	.92	-	.92	.93	.94	.89	.81	.79	.85	.86	-

(Erklärungen: HH=Hamburg, G=Göttingen, Tü=Tübingen, Ffm=Frankfurt/Main, Mü=München, Lüb=Lübeck, HD=Heidelberg, Be=Bern, Chic=Chicago, E_s=Edmonton Seniors, L=London, K=Kampala, E_i=Edmonton Juniors)

Berechnet man auf der Basis dieser experimentell gefundenen Werte die durchschnittlichen Reliabilitätskoeffizienten, so ergeben sich für Form A ein $r = .89$ und für Form B ein $r = .90$. Der Betrachter der Tabelle mag dabei zu dem Eindruck gelangen, der Test habe sich in Experimenten mit ausländischen Studenten als weniger reliabel erwiesen. Die niedrigeren Reliabilitätskoeffizienten auf der rechten Seite der Tabelle finden jedoch ihre Erklärung in der größeren Homogenität des Leistungsstands bei den ausländischen Studentengruppen, die eine geringere Standardabweichung der Meßwerte zur Folge hat. Je geringer jedoch die Standardabweichung einer Messung ist, desto niedriger wird auch deren Reliabilität (9).

3.4 Gültigkeit des Tripod Test

Die Qualitätskontrolle ist, wie bereits erwähnt, erst vollständig mit einer Gültigkeitsüberprüfung. Für den Tripod Test wurden mehrere Verfahren angewandt.

3.4.1 Kriterienbezogene Gültigkeit

Außenkriterien für die in der untenstehenden Tabelle wiedergegebenen Gültigkeitskoeffizienten waren die Examensleistungen der Studenten, die kurz vor diesen Prüfungen den Tripod Test bearbeitet hatten. Es handelt sich bei den Examina z. T. um schriftliche, z. T. um mündliche Prüfungen wie sie bei uns bislang üblich waren. Im Falle der Universität Kampala handelte es sich um eine kombiniert mündlich/schriftliche Prüfung (9).

Tabelle 2: Gültigkeitskoeffizienten des Tripod Test, ermittelt aus den Korrelationen von Testleistungen mit Leistungen in mündlichen und/oder schriftlichen Examina

	schriftlich		mündlich			kombiniert
	Chic	E	HD	Mü	Be	K
Form A	.78	.70	.71	.67	.71	.65
Form B	.74	-	.67	.70	.72	-

(Erklärungen siehe Tabelle 1)

3.4.2 Inhaltliche Gültigkeit

In Begleitung der Erprobungen des Tripod Test mit Studenten wurden von 60 Professoren und Hochschullehrern aus allen beteiligten Ländern Bewertungen hinsichtlich der Wichtigkeit der Fragen für eine ärztliche Prüfung erhoben. Die Fragen waren auf eine 4-Punkte-Skala einzuordnen, deren Werte definiert waren als: 1 = wichtig; 2 = eher wichtig als unwichtig; 3 = eher unwichtig als wichtig; 4 = unwichtig. Die Durchschnitte der Bewertungen pro Frage wurden für englisch- und deutschsprachige Professoren getrennt und zusammengefaßt berechnet. Die folgende Tabelle gibt die durchschnittlichen Bewertungen der beiden Testformen durch die beiden Gutachtergruppen wieder.

Tabelle 3: Mittlere Bewertung des Tripod Test durch deutsche und englische Professoren

	Test A	Test B
Professoren deutsch	1,58	1,53
Professoren englisch	1,66	1,80

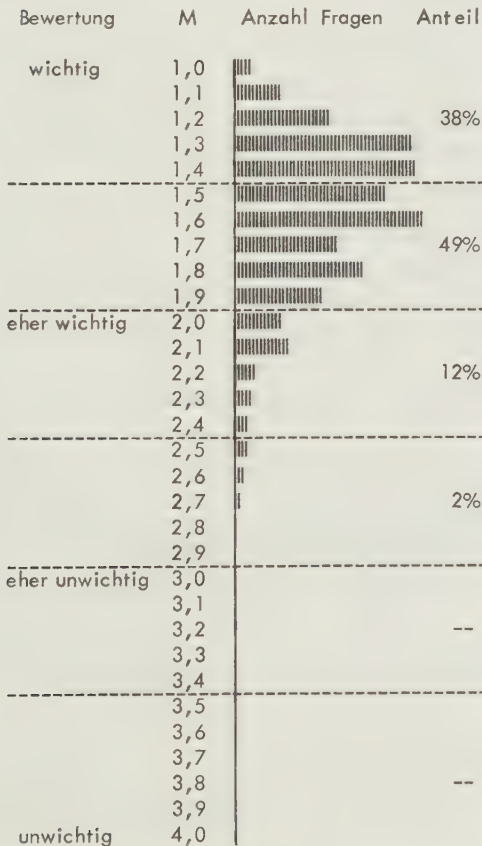
Dabei fällt vor allem für Test B eine signifikant schlechtere Bewertung durch die englischsprachigen Gutachter auf. Als Hypothese für eine Erklärung dieser Differenz wurde von Kapuste (9) ein deutsches "bias", d.h. eine mögliche Benachteiligung der englischsprachigen Studenten durch den von deutschen Ärzten konstruierten und redigierten Test angenommen, das von den englischsprechenden Gutachtern erkannt und entsprechend bewertet worden sein könnte. Diese Hypothese ließ sich jedoch nach verschiedenen Vergleichen von Leistungen der in- und ausländischen Studenten nicht aufrecht erhalten, so daß Qualitätsunterschiede zwischen der deutschen und englischen Form unwahrscheinlich sind.

Eine mögliche Erklärung der Bewertungsunterschiede liegt jedoch in der unterschiedlichen Darbietungsform des Fragenmaterials an die Gutachter. Während im deutschen Sprachraum die Hochschullehrer entsprechend ihres Kompetenzbereichs ausgewählte Fragengruppen zur Beurteilung vorgelegt bekamen, wurde den englischen Professoren das gesamte Testmaterial vorgelegt. Man könnte daher vermuten, daß ein Ermüdungseffekt bei der Beurteilung von 360 Testfragen zu einer

schlechteren Bewertung der in der Reihenfolge späteren Fragen und damit in aller Regel besonders der Fragen des Tests B geführt haben könnte (9).

Faßt man indessen die Bewertungen der Gutachter aus beiden Sprachräumen zusammen, so ergibt sich für beide Testformen übereinstimmend ein recht positiver Mittelwert von 1,60. Die folgende Abbildung zeigt die Verteilung der durchschnittlichen Bewertung von 360 Tripod Testfragen zusammengefaßt für deutsche und englische Professoren.

Abbildung 1: Durchschnittliche Bewertung der 360 Tripod Testfragen durch deutsche und englische Professoren



3.4.3 Gültigkeit des Tests gemessen an dem verwendeten Außenkriterium und den Professorenbewertungen

Es zeigt sich also, daß der Test von den Experten im Durchschnitt für eher wichtig gehalten wurde. Außerdem weisen die Testergebnisse eine hohe Übereinstimmung mit den Beurteilungen in offiziellen Prüfungen auf. Insbesondere fällt die hohe Korrelation zwischen den mündlichen Prüfungen und dem Test auf (vgl. Tab. 2). Dies bedeutet, daß das Abschneiden von Studenten im Tripod Test eine recht zuverlässige Voraussage ihrer Leistungen in den entsprechenden Fächern im medizinischen Staatsexamen bisheriger Prägung zuließ, welches demnach offenbar selbst eine höhere Reliabilität hatte, als mancher ihm zugestehen mochte (4).

Dies wird noch eher erklärlich, wenn man sich klarmacht, daß es sich bei dem bisherigen Staatsexamen doch immerhin um eine "Testbatterie" von 17 Einzelprüfungen durch in der Mehrzahl erfahrene Prüfer gehandelt hat, aus deren Notendurchschnitt man wohl zurecht ein recht zuverlässiges (=reliables) Gesamturteil erwarten konnte.

Eine Voraussage der Eignung der Studenten für den ärztlichen Beruf bzw. der zu erwartenden Qualität ärztlicher Leistungen erlaubt der Test jedoch nur in dem Maße, als das Staatsexamen selbst eine berufszielbezogene Gültigkeit im einleitend dargestellten Sinne besitzt oder besaß.

Eine Gültigkeitskontrolle des Staatsexamens bisheriger Prägung konnte und sollte nicht Aufgabe der vorliegenden Arbeit sein. Wohl aber konnte versucht werden, den Tripod Test durch eine weitere berufserfahrene Sachverständigengruppe auf seine inhaltliche Gültigkeit für die ärztliche Berufspraxis überprüfen zu lassen. Über ein solches Experiment soll im folgenden berichtet werden.

4. PROBLEMSTELLUNG ZUM EXPERIMENT

4.1 Auswahl der Sachverständigen

Bei der Auswahl der Sachverständigen trat ein erstes Problem auf, das in der fehlenden Definition des Berufs- bzw. Ausbildungszieles seine Ursache hat. Der Begriff, der in § 3 der Bestallungsordnung in Verbindung mit der Formulierung des Ausbildungszieles verwendet wurde und der in der Einführung zur neuen Approbationsordnung wieder auftaucht, ist der des "Arztes". Dies verschafft wenig Klarheit über das Berufsziel, da bei dem gegenwärtigen Stand der Spezialisierung und Arbeitsaufteilung im Bereich ärztlicher Tätigkeit das Prädikat Arzt nurmehr den Charakter eines Oberbegriffs hat, und sich unter "ärztlicher Beruf" eine Vielzahl von Tätigkeitsbereichen subsumieren läßt, die von einem einzelnen "Arzt" nicht mehr alle ausgefüllt werden können. Aus der Praxis weiß man jedoch, daß der gemeinte Bereich ärztlicher Tätigkeit derjenige des praktischen Arztes ist.

Von welcher Gruppe von Sachverständigen war also am ehesten zu erwarten, daß sie eine einheitliche und realistische Vorstellung von dem zu erstrebenden Ziel einer medizinischen Ausbildung habe? Es liegt nahe, die realistischsten Vorstellungen von diesem Arbeitsbereich bei denen zu erwarten, die in ihm tätig sind. Die kompetentesten Sachverständigen für die Kontrolle der berufszielbezogenen Gültigkeit eines medizinischen Leistungstests waren also in den Reihen praktischer Ärzte zu suchen.

Allerdings war zu berücksichtigen, daß auch der Tätigkeitsbereich praktischer Ärzte noch immer recht unterschiedlich sein kann, je nach Standort und Ausrüstung einer Praxis sowie den individuellen Neigungen des einzelnen Praktikers. So wäre beispielsweise nicht auszuschließen, daß, auch bei längerer praktischer Erfahrung, ein Landarzt bestimmte für den Stadtarzt spezifische diagnostische oder therapeutische Probleme nicht kennenlernt und umgekehrt (5).

Aus diesem Grunde war als zusätzliche Qualifizierung zu der praktischen Erfahrung von den Sachverständigen zu fordern, daß sie über ihren persönlichen Arbeitsbereich hinaus zumindest theoretisch eine

gute Kenntnis aller von ihnen und ihren Kollegen abzudeckenden Bereiche praktisch-ärztlicher Versorgung besäßen. Eine Gruppe von Ärzten, die sich um die Abgrenzung und Beschreibung dieser Bereiche sowie um deren stärkere Berücksichtigung innerhalb der medizinischen Ausbildung bemüht, sind die in den nationalen und in der Internationalen Gesellschaft für Allgemeinmedizin (IGAM) organisierten Allgemeinärzte. Durch die Kooperation von Mitgliedern dieser Organisation mit dem Institut für Ausbildungsforschung gelang es, 47 erfahrene Praktiker, die zusätzlich über berufstheoretische Kenntnisse verfügten, für eine erneute Überprüfung der Gültigkeit des Tripod Tests zu gewinnen.

4.2 Durchführbarkeit der Befragung

Ein weiteres Problem war der enorme Zeitaufwand, den die Beurteilung des Tripod Tests beanspruchen würde. Wenn die Zeit für die Beantwortung der 180 Fragen eines Testheftes durchschnittlich 3 Stunden beträgt, so mußte für eine sorgfältige Prüfung und Bewertung der Fragen mindestens ebensoviel Zeit veranschlagt werden. Einen solchen Zeitaufwand aber schien einem Allgemeinarzt, der durch seine Praxisarbeit in der Regel ohnehin überlastet ist, kaum zuzumuten zu sein. Eine Möglichkeit dieser Schwierigkeit zu begegnen, hätte in einer weiteren Aufteilung der Testfragen auf kleinere Pakete von Testfragen bestehen können, um so die Bearbeitungszeit zu verkürzen. Hierauf mußte jedoch vor allem deshalb verzichtet werden, weil damit die Zahl der notwendigen Sachverständigen über das mögliche Maß angestiegen wäre. Es war also fraglich, ob eine ausreichende Zahl qualifizierter Sachverständiger bereit sein würde, den notwendigen Zeitaufwand für die Beurteilung der Testfragen aufzubringen. Andererseits schien ein begrenzter Optimismus diesbezüglich gerechtfertigt, da das berufstheoretische Engagement der zu befragenden Ärzte ein überdurchschnittliches Interesse an der Frage nach der berufszielbezogenen Gültigkeit von Prüfungsfragen erwarten ließ.

4.3 Zuverlässigkeit der Ergebnisse

Die Zuverlässigkeit mit der sich aus den Beurteilungen Aussagen über die Gültigkeit der untersuchten Fragen ableiten ließen, würden wesentlich abhängen von der tatsächlichen Kompetenz und Einigkeit der Sachverständigen. Die Kriterien, die unter dem Aspekt der Kompetenz zur Auswahl einer ganz bestimmten Gruppe von Ärzten geführt hatten, wurden bereits genannt: Praktische und berufstheoretische Erfahrung. Sie ließen zugleich eine höhere Einigkeit erhoffen, wenn man davon ausging, daß die Ärzte sich in ihrem Urteil nicht nur auf ihre persönlichen Erfahrungen beziehen würden, sondern ein Berufsbild im Auge haben würden, das, erweitert durch Erfahrungsaustausch und Diskussion mit Fachkollegen aus der Allgemeinmedizin, ein hohes Maß von Allgemeingültigkeit beanspruchen könnte. Unter diesen Voraussetzungen waren auch von einer relativ kleinen Gruppe von ca. 20 bis 25 Experten für jede der beiden Testformen verlässliche und eindeutig interpretierbare Ergebnisse zu erwarten. Hierfür sprachen auch die positiven Erfahrungen hinsichtlich der Zuverlässigkeit, die bei der Begutachtung der Fragen durch Professoren gewonnen worden waren. Auch hier lagen die Gutachterzahlen pro Frage und Test in diesem Größenordnungsbereich (9).

4.4 Welche neuen Ergebnisse sind zu erwarten

Folgende Hypothesen bestanden bezüglich der aus der erneuten Gültigkeitsprüfung zu erwartenden Ergebnisse. Grundsätzlich wurde erwartet, daß die Beurteilungen der Allgemeinpraktiker deutliche Verschiedenheiten von denen der früher befragten Experten aus dem Bereich der Lehre und Forschung aufweisen würden. Und zwar wurde eine insgesamt weniger positive Bewertung erwartet. Es wurde angenommen, daß die Praktiker im allgemeinen nur solchen Fragen eine positive Bewertung geben würden, die einen unmittelbaren Bezug zu ihrer Tätigkeit haben, und andere als entweder theoretisch-akademisch oder dem klinischen Bereich zugehörig eher ablehnen würden. Das heißt, es wurde weniger eine allgemein negativere Bewertung als durch die hochschulischen Experten erwartet als vielmehr eine deutliche Trennung zwischen wichtigen positiv bewerteten einerseits und praktisch belang-

losen und daher negativ bewerteten Fragen andererseits erhofft. Es sollte außerdem geprüft werden, ob positive und negative Bewertungen von Fragen im systematischen Zusammenhang gebracht werden konnten mit möglichen Einteilungen der Fragen nach verschiedenen Einteilungskategorien; ob z.B. Fragen aus der Inneren Medizin signifikant positiver abschnitten, als solche aus der Pathologie oder solche aus dem Gebiet der Infektionskrankheiten positiver als Fragen aus der Endokrinologie. Im einzelnen wurden zahlreiche Einteilungskategorien verwendet, auf die weiter unten noch näher eingegangen werden soll.

5. Methodik

Durch drei deutsche Mitglieder der Internationalen Gesellschaft für Allgemeinmedizin wurden dem Institut für Ausbildungsforschung 75 in der BRD niedergelassene Allgemeinärzte als kompetente Sachverständige im Sinne der genannten Definitionen benannt. Da diese Ärzte über das ganze Bundesgebiet verstreut praktizieren, wurde eine schriftliche Befragung auf postalischem Wege vorgenommen.

5.1 Befragungsmodus

Jeder der 75 befragten Ärzte erhielt ein Exemplar des Tripod Test (Form A oder B) mit einem vorbereiteten Bogen für die Bewertungen der Fragen sowie zwei Begleitschreiben und Informationsmaterial über den Tripod Test. Die Verteilung der Testformen A bzw. B an die Ärzte erfolgte nach einem Zufallsverfahren.

Der Bewertungsbogen, der den Testheften beigegeben war, enthielt in übersichtlicher Gliederung eine 11-Punkte Skala für jede der 180 Fragen. In einem Begleitschreiben war der folgende Text als Anleitung für die Gutachten enthalten:

"Ziel des Projekts: Aus einem internationalen medizinischen Leistungstest, dessen Fragen von Universitätsprofessoren stammen (Tripod Test), sollen die Fragen bestimmt werden, die für die allgemeinärztliche Tätigkeit "gültig" sind. Als gültig gelten in der Testtheorie Fragen, für die gezeigt werden kann, daß sie tatsächlich messen, was gemessen werden soll. Gültig für die allgemeinärztliche Tätigkeit sind demnach

solche Fragen, die Kenntnisse oder Überlegungen verlangen, die ein Allgemeinpraktiker im Rahmen seiner ärztlichen Tätigkeit an irgendeiner Stelle benötigt. Wie man darlegen kann, gilt dies keineswegs für alle Kenntnisse, die von meist als Kliniker oder Theoretiker in hohem Grade spezialisierten Universitätslehrern gelehrt und geprüft werden.

Methode: Allgemeinpraktiker, die sowohl Erfahrung in eigener Praxis haben, als auch die Aufgaben anderer Allgemeinpraktiker (sei es auf dem Land oder sonst unter anderen Bedingungen) kennen, werden gebeten, die Prüfungsfragen nach diesen Erfahrungen zu beurteilen und auf einer Skala einzustufen. Die Skala hat folgende Werte:

gänzlich unwichtig	0	50	100	außerordentlich wichtig für die All- gemeinpraxis
für die Allgemeinpraxis	0 0 0	0 0 0 0 0	0 0 0	

Die Bedeutung der Werte 0 und 100 wurden oben formuliert, Zwischenwerte sind nach Ermessen zu wählen."

Im zweiten Begleitschreiben richteten sich die drei Ärzte, welche dem Institut die Adressen vermittelt hatten, an ihre Kollegen. Sie brachten darin u. a. zum Ausdruck, daß sie das Vorhaben für sinnvoll und wichtig hielten. Deshalb baten sie ihre Kollegen, sich auch durch den relativ großen Arbeitsaufwand nicht abschrecken zu lassen und das Vorhaben durch ihre Mitarbeit aktiv zu unterstützen. Dieses Schreiben sollte zur Stärkung der Motivation aus kollegialer Unterstützungsbereitschaft neben wissenschaftlichem Interesse beitragen.

5.2 Auswertung der erhaltenen Daten

5.2.1 Berechnung der durchschnittlichen Bewertung der einzelnen

Fragen und der gesamten Tests aus den Mittelwerten der Gutachterurteile

Bei der statistischen Auswertung der Daten wurde zunächst mit Hilfe des Computerprogramms BMD 07D (3) für jede der 360 Fragen die durchschnittliche Bewertung durch die praktischen Ärzte sowie die Standardabweichung der zugrundeliegenden Einzelurteile errechnet.

Ferner lieferte das Programm für jede Aufgabe eine graphische Darstellung der Verteilung der Gutachterurteile auf die 11 Skalenpunkte.

In einer Graphik wurde mit Hilfe der Parameter Mittelwert und Standardabweichung die Verteilung von 180 Fragen des Test A bzw. B in einem Koordinatensystem dargestellt.

Auf der Basis der durchschnittlichen Bewertung der einzelnen Fragen wurden mit Hilfe des Computerprogramms BMD 02 D (3) Gesamtdurchschnitte und Standardabweichungen für beide Testformen berechnet.

2.2.2 Berechnung von Durchschnitt und Verteilung der Bewertungen bei den einzelnen Gutachtern

Mit einem speziellen Zählprogramm wurde für jeden Gutachter der Durchschnitt der von ihm vergebenen Wertungen und deren Standardabweichung berechnet und ebenfalls eine graphische Darstellung der Verteilung seiner Urteile über die vorgegebene Skale ausgedruckt.

Mittelwert und Standardabweichung der Gutachter dienten wiederum zur graphischen Verdeutlichung der Verteilung der Durchschnittswerte in einem Koordinatensystem.

5.2.3 Berechnung der Korrelationen der gewonnenen Gutachterurteile mit den vorhandenen Professorenurteilen, sowie der Urteile der Gutachter untereinander

Um festzustellen, inwieweit die Urteile der Praktiker mit denen der Hochschullehrer übereinstimmten, wurden die Korrelationen der Durchschnitte beider Gutachterstichproben miteinander berechnet. Eine solche Korrelation ist als Außenkriterium für die Ärzteurteile zu werten.

Um ein Kriterium für die Zuverlässigkeit der Gutachterurteile zu gewinnen, wurden die Korrelationen der Ärzte in ihren Urteilen miteinander berechnet. Außerdem wurde eine Faktorenanalyse dieser Korrelationen durchgeführt, um Anhaltspunkte für hypothetisch mögliche systematische Auffassungsunterschiede zwischen verschiedenen Praktikergruppen zu erhalten.

5.2.4 Überprüfung von Zusammenhängen zwischen der durchschnittlichen Bewertung einzelner Fragengruppen ("Subtests") und ihrer Zuordenbarkeit zu verschiedenen inhaltlichen und formalen Kategorien

Eine weitere statistische Analyse des Datenmaterials hatten zum Ziel, Gesetzmäßigkeiten im Urteilsverhalten der Gutachter aufzufinden. Hier-

zu wurden Fragen der Tests nach verschiedenen inhaltlichen und formalen Kriterien in Gruppen zusammengefaßt und überprüft, ob sich signifikante Häufungen positiver oder negativer Urteile in einzelnen Fragengruppen finden ließen. Dabei wurden folgende Kategorisierungen angewandt (9):

- Inhaltliche Kategorien:
1. Herkunft der Fragen nach Fächern (Innere Medizin, Chirurgie etc.)
 2. "Querverbindung" (Infektionskrankheiten, Kardiologie etc.)
 3. Ärztlicher Anwendungsbereich (Therapie, Diagnose etc.)
 4. Ätiologie (Neoplasma, Kongenitale Anomalie etc.)
- Formale Kategorien:
5. Frageform "echter Fall"
 6. Teststatistische Parameter (Trennschärfe, Prozent richtige Antworten)

Bei der Einteilung der Fragen nach Fächern handelte es sich um den Fachbereich, in dem der Hochschullehrer, der die betreffende Frage vorgeschlagen hatte, tätig ist. Aus diesem Grunde treten in einzelnen Fällen Überschneidungen auf, die eine Zuordnung zu einem Fachbereich unbefriedigend erscheinen lassen.

Die übrigen Inhaltskategorisierungen waren von den Herausgebern des Tests (Götz, Noack) vorgenommen worden, um eine differenzierte Analyse der Experimente mit dem Tripod Test zu ermöglichen. Ebenso auch die Zuordnung zu der formalen Kategorie "echter Fall", in der solche Fragen enthalten sind, welche die Schilderung einer tatsächlichen klinischen Situation bzw. eines Patienten und seiner Beschwerden oder Befunde zur Grundlage haben.

Die teststatistischen Kennwerte Schwierigkeit und Trennschärfe der Fragen wurden aus Experimenten mit 1330 deutschen und 410 englischen Studenten gewonnen, über die bereits weiter oben berichtet wurde (s. S.13). Die Kategorie Fragenart berücksichtigt die Tatsache, daß der Test drei verschiedene Fragentypen enthält, neben den gewöhnlichen MCF noch "Weil-Fragen" und "Oberbegriff-Fragen". Beispiele hierfür werden im Ergebnisteil angeführt (9).

Das Verfahren, mit dem die oben erwähnte Fragestellung untersucht werden sollte, bestand in der Ermittlung der Zufallswahrscheinlichkeit der Verteilung der Fragen in Mehrfeldertafeln, in die einerseits die kategorielle Zuordnung der Fragen, andererseits ihre Bewertung durch die ärztlichen Gutachter einging. Die folgende Tabelle 4 zeigt zur Veranschaulichung das Beispiel eines solchen Mehrfeldertests, in dem der Einfluß der Zugehörigkeit zu der Kategorie "Infektionskrankheiten - Chemotherapie" auf die Beurteilung der Fragen durch die Gutachter untersucht wird.

Die durchschnittlichen Bewertungen der Fragen in dem Bereich von 0 - 100, der in der verwendeten 11-Punkte-Skala vorgegeben war, wurden für die Mehrfeldertests der Übersichtlichkeit halber in vier Bewertungsklassen zusammengefaßt. Die Zuordnung der Bewertungsdurchschnitte zu den Bewertungsklassen ist in der Tabelle 5 wiedergegeben.

Tabelle 4: Bewertung von Fragen des Subtests "Infektionskrankheiten - Chemotherapie" im Vergleich zu den übrigen Fragen des Tests

Bewertungsklasse	1	2	3	4	Anzahl
Infektionskrankheiten	9	4	0	0	13
andere	54	76	28	9	167
Summe	63	80	28	9	180
chi-Quadrat: 8,16	Rpbi: - .20		M: 1,31		

Tabelle 5: Zusammenfassung der Praktikermittelwerte in vier Bewertungsklassen

Mittelwert		Bewertungsklasse
0 - 24	=	4
25 - 49	=	3
50 - 74	=	2
75 - 100	=	1

Das angewandte Computerprogramm (BMD 08 D) (Dixon 1967) lieferte neben der Mehrfeldertafel selbst den chi-Quadratwert der Verteilung als Maß der Zufallswahrscheinlichkeit der gefundenen Verteilung, sowie deren punktbiserialen Korrelationskoeffizienten R_{pb} , aus dessen Vorzeichen sich die Tendenz der Verteilung - zu eher positiven oder negativen Bewertungen - in der betreffenden Kategorie ablesen läßt. Zusätzlich wurde der Mittelwert M aus den Durchschnittswerten der Fragen der betreffenden Kategorie berechnet, um ihn mit dem Mittelwert des Gesamttests vergleichen zu können.

In dem gewählten Beispiel finden sich von den 13 Fragen der untersuchten Kategorie alle in den beiden oberen Bewertungsklassen. Die Zufallswahrscheinlichkeit eines solchen Befundes liegt unter 5 %, was sich in dem signifikanten chi-Quadratwert ausdrückt ($\chi^2_{DF3/5\%} = 7,82$) (12). Außerdem handelt es sich um eine mathematisch negative Korrelation, da mit numerisch steigendem Mittelwert die Zahl der Fragen, die der Kategorie zugehören, abnimmt. Im Sinne der Fragestellung handelt es sich jedoch um eine inhaltlich positive Korrelation, weil die gute Bewertung einem numerisch niedrigen Mittelwert entspricht. Der Mittelwert dieses Subtests beträgt 1.31 gegenüber $M = 1.91$ für den Gesamttest.

6. Ergebnisse

6.1 Rücklauf

Insgesamt 50 Ärzte sandten den Test bewertet zurück. Das entspricht einer Rücklaufquote von 66%. Zeitlich verteilte sich der Rücklauf wie folgt auf die einzelnen Phasen der Befragung: 30 ausgefüllte Bewertungsbogen gingen innerhalb der ersten vier Wochen ein. Auf ein erstes Erinnerungsschreiben hin gingen 6 weitere Beurteilungen ein. Durch telefonische Erinnerung nach 4 weiteren Wochen konnte schließlich noch in 14 Fällen eine Bearbeitung des Tests erreicht werden, von denen allerdings drei erst nach einem vereinbarten Zeitlimit eintrafen und aus organisatorischen Gründen in die statistische Auswertung

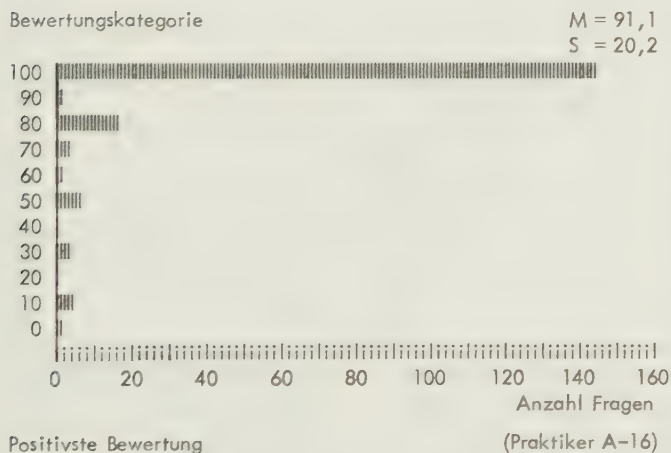
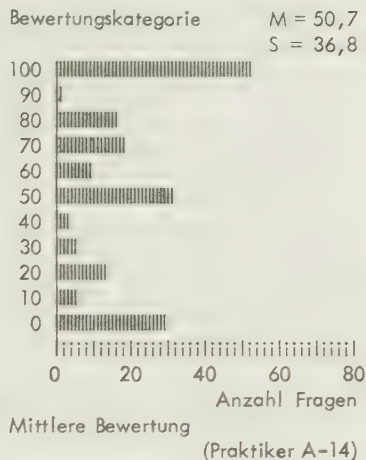
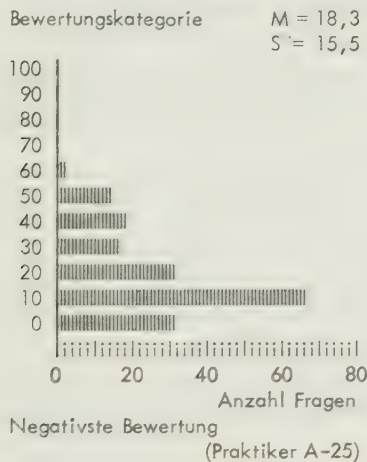
nicht mehr einbezogen werden konnten. Ausgewertet wurden 24 Bewertungen des Test A und 23 Bewertungen des Test B. Die Beurteilungsbogen enthielten bei 180 Fragen im Durchschnitt 178,6 Bewertungen, wurden also praktisch immer vollständig bearbeitet.

6.2 Ergebnisse der Auswertung

Durchschnitt und Verteilung der Bewertungen bei einzelnen Gutachtern

Die durchschnittliche Bewertung des gesamten Tests durch einzelne Praktiker schwankt in weiten Grenzen. Das negativste Durchschnittsurteil eines Gutachters lautete 18.8 für Test A und 37.7 für Test B. Die positivsten Bewertungen liegen bei 92.1 (Test A) bzw. 90.1 (Test B). Die Standardabweichungen der Verteilungen schwanken in einem Bereich von 16.0 bis 43.4 in Test A und von 21.2 bis 38.5 in Test B. Drei Beispiele für extreme Verteilungen der Bewertungen bei einzelnen Gutachtern zeigt die folgende Abbildung:

Abbildung 2: Beispiele für die Verteilung der Einzelurteile über die Skala bei drei Gutachtern



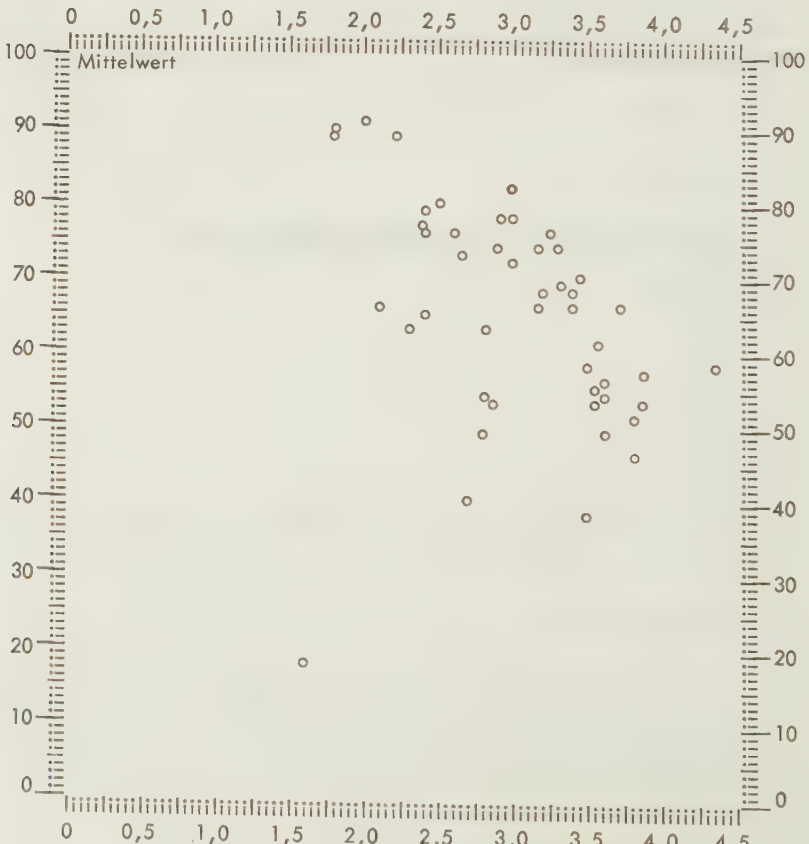
Bei dem ersten Beispiel handelt es sich um die negativste Bewertung die der Test erfuhr. Der betreffende Gutachter verwendete fast ausschließlich die untere Hälfte der Skala, hielt also die überwiegende Zahl der Fragen für eher unwichtig oder sogar gänzlich unwichtig.

Das entgegengesetzte Extrem zeigt das dritte Beispiel. Dieser Gutachter hielt allein 144 Fragen für außerordentlich wichtig. Immerhin benutzte dieser Gutachter einige Male auch die negative Hälfte der Skala, wenn auch nur in wenigen Fällen.

Dazwischen liegt das Beispiel der Abbildung. Dieser Gutachter schöpfte die Bewertungsmöglichkeiten der Skala voll aus und fällte annähernd gleich viele positive wie negative Urteile.

Die folgende Graphik zeigt die zweidimensionale Verteilung der Mittelwerte und Standardabweichungen für alle Gutachter.

Abbildung 3: Verteilung der 47 Gutachter nach den Parametern Mittelwert und Standardabweichung (Jeder Kreis repräsentiert einen Gutachter)



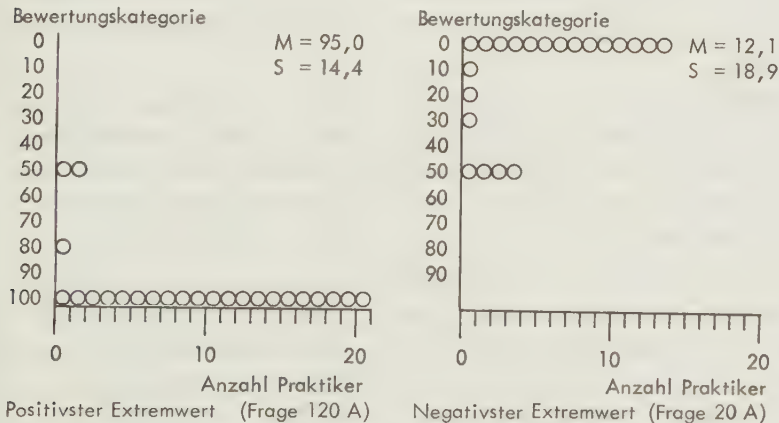
Die Darstellung zeigt für die Gutachterdurchschnittswerte eine negative Korrelation zwischen M und S. Die Koeffizienten dieser Korrelation betragen $-.58$ für Test A und $-.50$ für Test B. Weiterhin ist der Darstellung zu entnehmen, daß es sich in dem Fall des Gutachters mit der negativsten Bewertung um eine Handhabung der Bewertungsskalen handelt, die weit aus dem Rahmen der übrigen fällt.

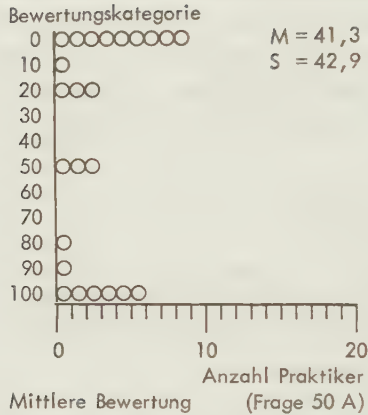
6.2.2 Durchschnittliche Bewertung der einzelnen Fragen und der gesamten Tests

Die Berechnung der Bewertungsdurchschnitte für die 2 mal 180 Fragen erbrachte Werte, die von 12.1 bis 96.7 in Test A und von 16.8 bis 94.4 in Test B reichten. Die Standardabweichungen der Bewertungen pro Frage schwankten in einem Bereich von 11.8 bis 43.0 bzw. 10.4 bis 41.7 in Test A bzw. Test B.

Die folgende Abbildung zeigt drei Extrembeispiele, wie Fragen bewertet wurden.

Abbildung 4: Drei Beispiele unterschiedlich bewerteter Fragen





Das erste Beispiel zeigt eine Frage, die von der überwiegenden Zahl der Ärzte die optimale Bewertung erhielt. Nur 3 der 24 Gutachter, die den Test bewertet hatten, vergaben ein anderslautendes Urteil. Auf diese Weise ergibt sich ein sehr hoher Mittelwert bei niedriger Standardabweichung, die das Maß der Streuung der Urteile über die Skala angibt.

Ebenfalls niedrig ist die Standardabweichung im 2. Beispiel. Hier handelt es sich jedoch um eine Frage, die von fast allen Gutachtern eher für unwichtig gehalten wurde. 14 der 24 Gutachter stufen sie auf den niedrigsten Skalenwert ein, so daß der sehr niedrige Mittelwert von 12.1 resultiert.

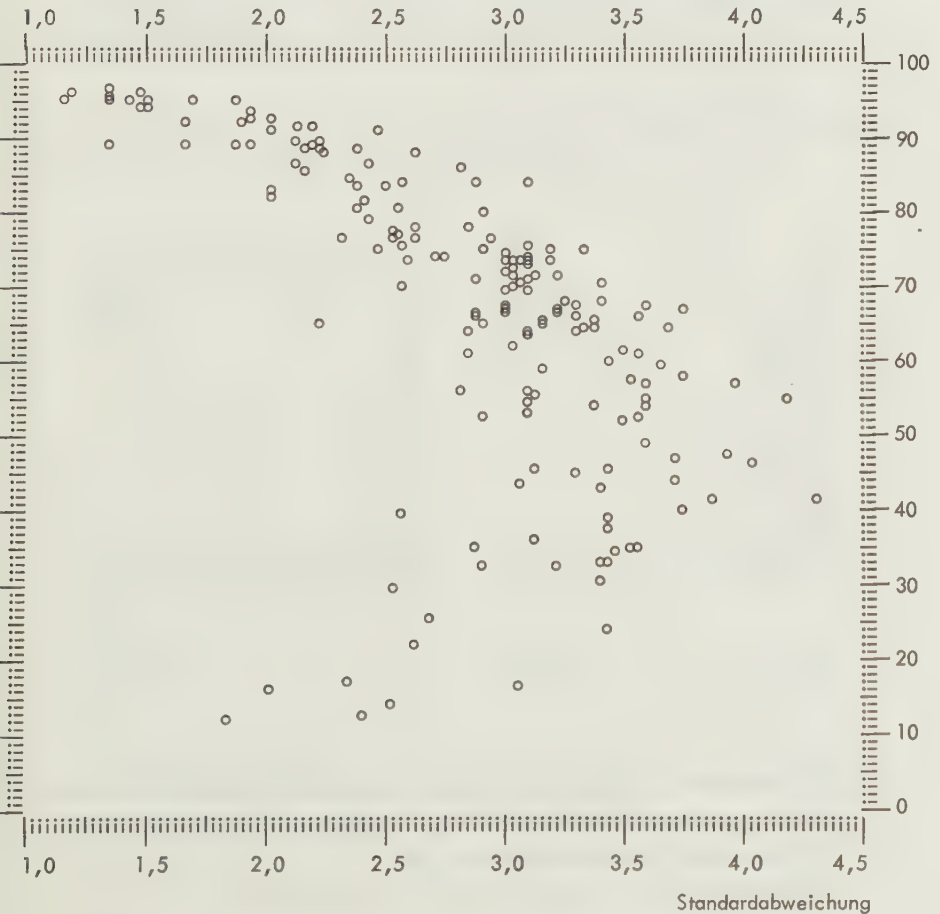
Im dritten Beispiel schließlich handelt es sich um eine Frage, die sehr unterschiedlich bewertet wurde. Die breite Streuung der Urteile, die von "gänzlich unwichtig" bis "außerordentlich wichtig" reichen, finden in einer hohen Standardabweichung ihren Niederschlag.

Um nun anschaulich zu machen, wie diese Bewertungstypen innerhalb der Gesamtheit der Fragenbewertungen einzuordnen sind, wird in der folgenden Abbildung die Verteilung aller Fragen nach den Parametern Mittelwert und Standardabweichung in einem Koordinatensystem dar-

gestellt.

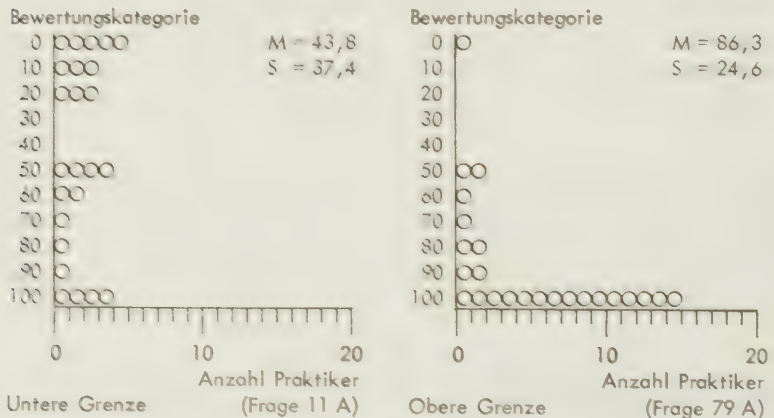
Für die Darstellung wurden der Übersichtlichkeit halber nur die Ergebnisse für die Fragen des Test A dargestellt. Diejenigen des Test B sind in allen wesentlichen Punkten gleich, sodaß sich eine gesonderte Darstellung erübrigt.

Abbildung 5: Verteilung von 180 Fragen des Tripod Test (Form A) nach den Parametern Mittelwert und Standardabweichung (Jeder Kreis repräsentiert eine Frage)



Dabei fällt in der gewählten Darstellungsform die Form der Verteilung auf. Sie spiegelt den soeben an drei Beispielen dargestellten Befund wider, daß prinzipiell drei Varianten von Fragenbewertungen auftreten, die in der gereinigten Darstellung die äußersten Schenkelpunkte bzw. den Scheitelpunkt der U-Verteilung einnehmen. Allerdings ist der Abbildung auch zu entnehmen, daß die gewählten Fragenbeispiele nicht als typisch für die durchschnittliche Fragenbewertung gelten können. Es sollen deshalb zwei weitere Bewertungsverteilungen von Fragen gezeigt werden, die an den Grenzen des Bereichs liegen, in dem zwei Drittel der mittelmäßigen bewerteten Fragen zu finden sind. An der oberen Grenze dieses Bereichs wurde die Frage 108 A, an der unteren die Frage 11 A für die Darstellung gewählt.

Abbildung 6: Bewertungsverteilungen von zwei Fragen an der oberen und der unteren Grenze des Bereichs, in dem sich die mittleren zwei Drittel der Fragen befinden



Errechnet man aus den Bewertungsdurchschnitten der Fragen die durchschnittliche Gesamtbewertung für die beiden Testformen, so ergeben sich folgende Werte: Für Test A ergibt sich ein Mittelwert von 66,3 bei einer Standardabweichung von 20,8, für Test B beträgt der Mittelwert 64,8 bei einer Standardabweichung von 20,5.

6.2.3 Korrelation der Gutachterurteile mit Professorenurteilen sowie der Urteile der Gutachter untereinander

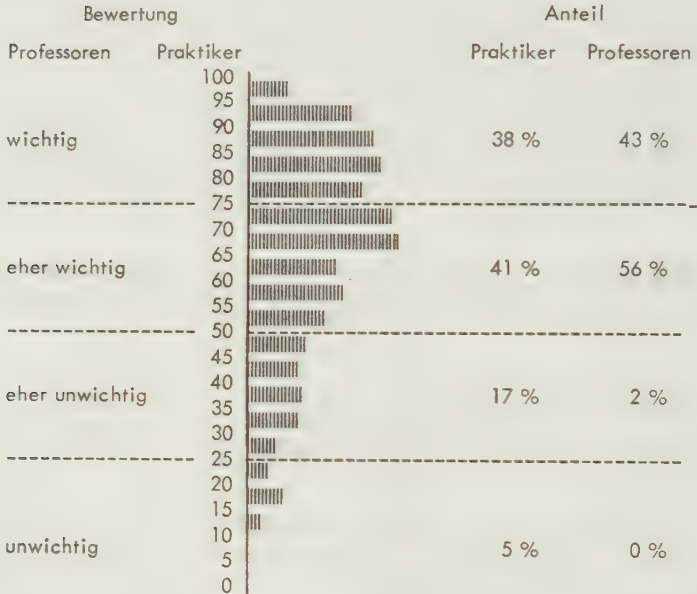
Die Berechnung der Korrelationen von Praktikerurteilen über die Testfragen mit denen von englischen bzw. deutschen Professoren erbrachte hohe positive Koeffizienten. Die Übereinstimmung in der Beurteilung der Testfragen zwischen den deutschen Ärzten und ihren deutschen Kollegen an den Hochschulen schlägt sich in Korrelationskoeffizienten von .66 für Test A und .75 für Test B nieder. Der Vergleich der Praktikerurteile mit denen englischer Hochschullehrer liegt niedriger mit .55 für Test A und .53 für Test B. Die folgende Tabelle faßt diese Ergebnisse zusammen.

Tabelle 6: Korrelationen zwischen Bewertungsdurchschnitten der Tripod Testfragen von Ärzten und Professoren

	Test A	Test B
Allgemeinpraktiker/deutsche Professoren	.66	.75
Allgemeinpraktiker/englische Professoren	.55	.53

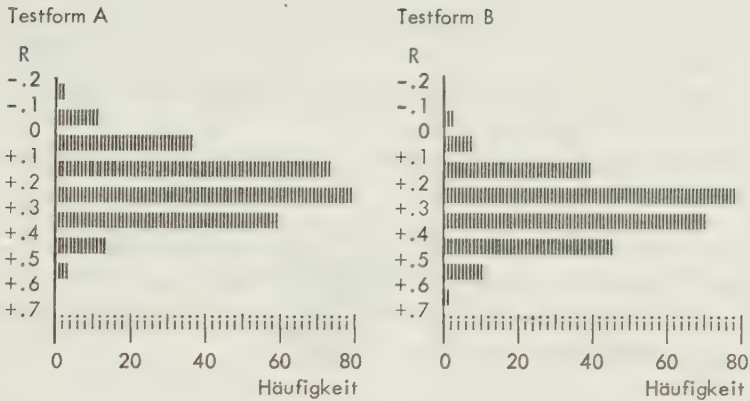
Der Vergleich der prozentualen Häufigkeit zeigt jedoch, daß der Test von den Praktikern insgesamt etwas schlechter als von den Professoren beurteilt wurde (vorausgesetzt, daß die Transformierung der Praktikermittelwerte keine wesentlichen Fehler in die Verteilung gebracht hat, was wir jedoch für unwahrscheinlich halten). So finden sich bei den Praktikern gut 20% der Fragen, die in die untere Hälfte der Skala fallen, während die entsprechenden Klassen bei den Professoren praktisch leer bleiben (Abb. 7).

Abbildung 7: Durchschnittliche Bewertung von 360 Tripod Testfragen durch deutsche Allgemeinpraktiker. Vergleich mit den zusammengefaßten Bewertungsdurchschnitten deutscher und englischer Professoren.



Als Maßstab für die Zuverlässigkeit der Urteile einzelner Ärzte innerhalb ihrer eigenen Gruppe wurde für jeden Praktiker die Korrelation seiner Bewertungen mit denen sämtlicher Kollegen, die den gleichen Test bewertet hatten, ermittelt. Das Maximum der Häufigkeit aller so ermittelten Korrelationskoeffizienten liegt für beide Tests im Bereich von .20 bis .29, der Durchschnitt bei .25. Im einzelnen liegen die höchsten zwischen zwei Praktikern erreichten Korrelationen im Bereich von .50 bis .59 für Test A bzw. .60 bis .69 bei Test B. Zufalls-Korrelationen sind demgegenüber nullverteilt.

Abbildung 8: Häufigkeit der gefundenen Korrelationen der Praktischen Ärzte untereinander



Eine Faktorenanalyse der Korrelationen ergab nur einen allgemeinen Faktor, was darauf hinweist, daß alle Gutachter nach ähnlichen Gesichtspunkten urteilen (9).

6.2.4 Zusammenhänge zwischen der Bewertung von Fragen und ihre kategoriale Zuordnung zu verschiedenen Subtests

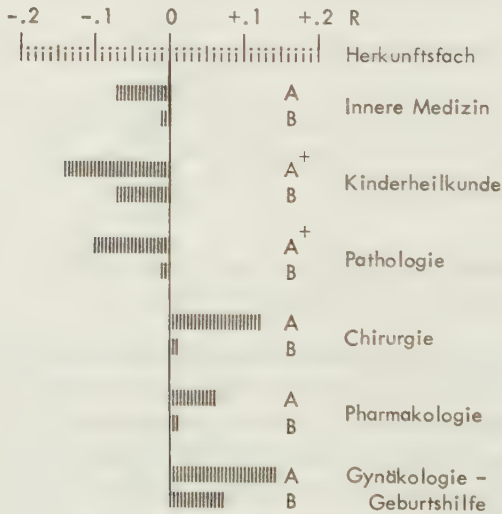
Die Suche nach Gesetzmäßigkeiten in der Beurteilung der Fragen im Sinne einer Bevorzugung oder Ablehnung bestimmter Gruppen von Fragen brachte statistisch signifikante Ergebnisse nur in einigen Fällen. Es erscheint jedoch gerechtfertigt und sinnvoll, auch solche Ergebnisse zu betrachten, deren Irrtumswahrscheinlichkeit größer als 5% ist, wenn eine entsprechende Tendenz in beiden Testformen das Vorhandensein von Zusammenhängen nahelegt oder aber, wenn deutlichere gegenläufige Tendenzen in beiden Tests zur Suche nach einer möglichen Erklärung solcher Befunde auffordern.

Mit den folgenden Abbildungen soll zunächst in einer graphischen Darstellungsform eine Übersicht über die wichtigsten angewandten Kate-

gorien und die gefundenen Bewertungstendenzen vermittelt werden. Zu diesem Zweck wurden die Korrelationskoeffizienten der Verteilung nach numerischem Wert und Vorzeichen als Kennwerte der jeweiligen Kategorien aufgetragen. Diejenigen Fälle, in denen die zugehörige Verteilung eine Irrtumswahrscheinlichkeit von weniger als 5% besitzt, wurden durch ein Pluszeichen hinter der Kennzeichnung der betreffenden Testform kenntlich gemacht.

Abbildung 9 zeigt die Korrelation der Praktikerbewertungen mit dem Herkunftsfach der Fragen.

Abbildung 9: Korrelation der Praktikerbewertungen mit dem Herkunftsfach (Ein Pluszeichen bei der Kennzeichnung der Testform bedeutet eine Irrtumswahrscheinlichkeit der betreffenden Korrelation von weniger als 5%)



In Abbildung 10 sind die Korrelationen der Praktikerbewertungen mit verschiedenen Querkategorien dargestellt, einer Kategorisierung, welche alle Fragen einer bestimmten systematischen Zugehörigkeit zusammenfaßt, unabhängig beispielsweise von ihrer Fachzugehörigkeit.

Abbildung 10: Korrelation der Praktikerbewertungen mit den Querkategorien

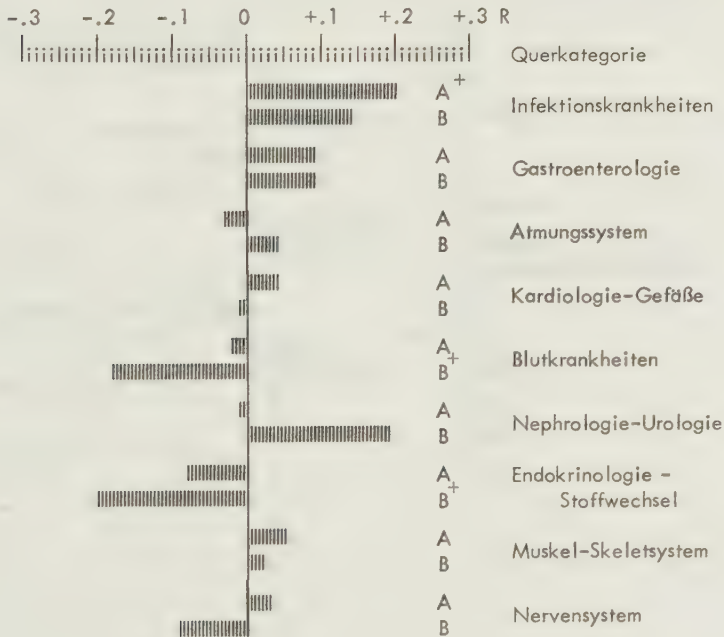


Abbildung 11 zeigt nun die Bewertungskorrelationen bei einer Einteilung der Testfragen nach Kriterien der Ätiologie der in ihnen behandelten Probleme.

Abbildung 11: Korrelation der Praktikerbewertung mit ätiologischen Kategorien



In Abbildung 12 sind die Korrelationen der Praktikerbewertungen mit Kategorien ärztlicher Anwendungsbereiche dargestellt.

Abbildung 12: Korrelation der Praktikerbewertungen mit ärztlichen Anwendungsbereichen

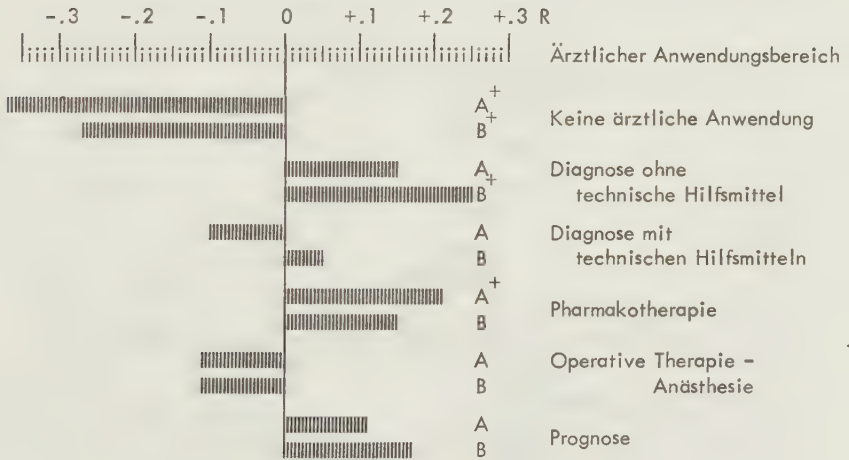
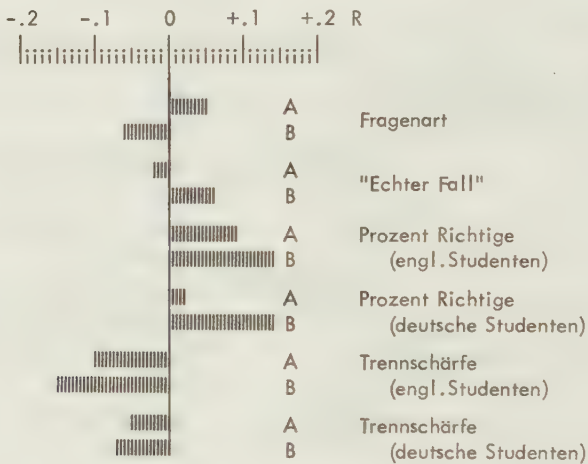


Abbildung 13 schließlich gibt die Korrelationen der Gutachterurteile mit formalen Kategorisierungen und teststatistischen Kennwerten der Fragen wieder.

Abbildung 13: Korrelation der Praktikerbewertungen mit Formalkategorien und statistischen Parametern



Im folgenden soll nun auf einzelne Befunde näher eingegangen und versucht werden, bestimmte Bewertungstendenzen anhand von Fragenbeispielen verständlich zu machen. Die bei den einzelnen Subtests angegebenen Mittelwerte sind auf der Basis der für diesen Teil der Untersuchung vorgenommenen Transgeneration der Praktikerurteile in vier Mittelwertsklassen zu verstehen, wie sie weiter oben beschrieben wurde (siehe Seite 26). Die auf dieser Basis errechneten Gesamtdurchschnittswerte betragen 1.91 für Test A und 1.89 für Test B. Bei der Auswahl der Beispiele wurde sowohl von dem objektiven Kriterium des Bewertungsdurchschnitts der Frage als auch vom subjektiven Urteil des Verfassers, inwieweit eine Frage als typisch für ihre Kategorie gelten konnte, ausgegangen.

6.2.4.1 Positiv bewertete Fragenkategorien

Eine der signifikant positiven Abhängigkeiten der Fragenbewertung von der kategoriellen Zuordnung findet sich in Test A für die Querverbindung "Infektionskrankheiten-Chemotherapie". Im Test B läßt sich die gleiche Tendenz ablesen, allerdings sind die Ergebnisse hier nicht signifikant auf dem 5% Niveau. Typische Beispiele aus dieser Kategorie sind die folgenden Aufgaben.

Fragenbeispiele aus den Subtests "Infektionskrankheiten - Chemotherapie"

Testform A

M= 1,31
N= 13

Testform B

M= 1,46
N= 13

128 A

- 1- Toxoid
- 2- Lebendvaccine und Gamma-globulin
- 3- orale Lebendvaccine
- 4- Totvaccine, parenteral
- 5- intradermale Lebendvaccine
- 6- Gammaglobuline
- 7- Antitoxin
- 8- keines der genannten

Durch welche der oben genannten Impfstoffe wird ein langfristiger Schutz gegen Poliomyelitis erreicht? (Wähle zwei)

170 A

Welche der folgenden Erkrankungen sind KEINE Indikationen für Penicillin G? (Wähle drei)

- 1- Streptokokkenpneumonie
- 2- Gonorrhoe
- 3- Diphtherie
- 4- Coliinfektion der Harnwege
- 5- Miliartuberkulose
- 6- Mykosen
- 7- Agranulocytose
- 8- Endocarditis lenta

126 B

Ein 7 Jahre altes Mädchen hat eine nachweisbar durch β -hämolytische Streptokokken bedingte Tonsillitis. Nach Angaben der Mutter habe eine frühere Gabe von Penicillin zu einer Urticaria geführt. Welche der folgenden Medikamente darf man dem Kind NICHT geben? (Wähle zwei)

- 1- Tetracycline für 10 Tage
- 2- Natriumoxacillin (z.B. Stapenor) für 10 Tage
- 3- Chloramphenicol bis der bakteriologische Befund negativ ist
- 4- Erythromycin für 10 Tage
- 5- Sulfisoxazol (Gantrisin) für 10 Tage
- 6- Penicillin und Benadryl (ein Phenothiazinderivat) oral für 10 Tage

Die positive Tendenz des Themenbereichs Infektionskrankheit auf der einen und Chemotherapie auf der anderen Seite schlägt sich noch in zwei weiteren Kategorisierungen nieder. So findet sich in der ätiologischen Kategorie "Infektion" eine Tendenz zu positiver Bewertung in beiden Tests, allerdings unterhalb der 5% Signifikanzgrenze. Zur Illustration dienen die Fragenbeispiele 125 B und 51 A aus dieser Kategorie.

Fragenbeispiele aus dem Subtest "Ätiologie-Infektion"

<u>Testform A</u>	M= 1.58 N= 19	<u>Testform B</u>	M= 1.74 N= 19
<u>51 A</u>		<u>125 B</u>	
1- Erysipel		Bei einem 2-jährigen Kind besteht	
2- Lobärpneumonie		seit 3 Tagen hohes Fieber und eine	
3- Meningitis		schmerzhafte Rötung und Schwellung	
4- Scharlach		im Bereich des rechten, unteren	
5- abszedierende Entzündung		Oberschenkels und Kniegelenks. Alle	
6- Endocarditis lenta		anderen Gelenke sind unauffällig und	
7- hämorrhagische Pneumonie		frei beweglich. Wichtige diagnostische	
8- Waterhouse-Friderichsen-Syndrom		Maßnahmen sind: (Wähle zwei)	
Welche der oben angeführten		1- Röntgenaufnahme	
Krankheiten ist in nahezu al-		2- Bestimmung der Leukozytenzahl	
len Fällen ihres Auftretens die		3- Bestimmung des Antistreptolysin-	
Komplikation einer Grippeer-		titers	
krankung? (Wähle eine)		4- Blutkultur	
		5- Tuberkulinprobe	
		6- Knochenmarkpunktion	

Noch deutlicher zeigt sich die positive Tendenz bei den Fragen der Anwendungskategorie "Pharmakotherapie". Hier liegt die Irrtumswahrscheinlichkeit der Verteilung der Fragen sogar unter 5% in Test A. Für Test B ist die Verteilung bei gleicher Tendenz nicht signifikant. Zwei Beispiele aus dieser Kategorie, die nicht zugleich den Infektionskrankheiten zuzurechnen sind, stellen die Fragen 158 A und 165 B dar.

Fragenbeispiele aus den Subtests Pharmakotherapie

Testform AM= 1.53
N= 32Testform BM= 1.59
N = 29158 A

Welche der folgenden Behauptungen sind richtig ?
(Wähle drei)

- 1- Natriumbromid ist bei Grand-mal-Epilepsie das Mittel der Wahl
- 2- Primidon (Mylepsin, Wz.) ist bei Grand-mal-Epilepsie wirksam
- 3- Diphenylhydantoin (Zentropil, Wz.) wird gegen Petit-mal-Anfälle mit großem Erfolg angewandt, obwohl es eine Gingivalhyperplasie hervorrufen kann
- 4- Trimethadion (Tridione, Wz.) wird gegen Petit-mal-Anfälle mit Erfolg angewandt, obwohl es eine aplastische Anämie hervorrufen kann
- 5- Bei Therapiebeginn werden Antiepileptika zunächst niedrig dosiert und die Dosen dann individuell gesteigert

165 B

Welche der folgenden Mittel bewirken bei längerer Gabe eine Schädigung des 8. Hirnnerven? (Wähle zwei)

- 1- Penicillin G
- 2- Tetracycline
- 3- Kanamycin
- 4- Polymyxin B
- 5- Erythromycin
- 6- Dihydrostreptomycin
- 7- Chloramphenicol
- 8- Sulfonamide

Bei der Bewertung der Fragen aus der Kategorie "Chemische Einwirkungen" findet sich ebenfalls eine geringe Tendenz zu positiver Einstufung, was mit dem Befund in der Kategorie "Anwendungsbereich: Pharmakotherapie" korrespondiert, da die Mehrzahl der Fragen zugleich in diese Kategorie gehört. Darüberhinaus finden sich Fragen aus dem engeren Bereich der Toxikologie, wie z.B. die Items 82 A oder 169 B.

Fragenbeispiele aus den Subtests "Chemische Einwirkungen"

Testform A

M= 1.80

N= 10

82 A

Welches ist die Primärreaktion der Blausäure im Organismus? (Wähle eine)

- 1- Bindung an Hämoglobin
- 2- Bindung an Hämoglobin
- 3- Erhöhung der Wasserstoffionen-konzentration
- 4- Hemmung der Katalase
- 5- Bindung an Cytochromoxydase

Testform B

M= 1.50

N= 6

169 B

Welche Symptome sind für die chronische Bleivergiftung typisch? (Wähle drei)

- 1- Hypotonie
- 2- periphere Neuritis mit Sensibilitätsausfällen
- 3- motorische Polyneuritis ohne Sensibilitätsausfälle
- 4- Anämie
- 5- Psychosen, Delirien
- 6- Schädigung der Nierentubuli

Weiterhin zeichnet sich eine überdurchschnittlich positive Bewertung für diejenigen Fragen ab, die Kenntnisse auf dem Gebiet der Diagnostik ohne technische Hilfsmittel erfordern. Die Zufallswahrscheinlichkeit der Verteilungen liegt allerdings nur für Test B unterhalb 5%.

Fragenbeispiele aus dem Subtest "Diagnose ohne technische Hilfsmittel"

Testform A

M= 1.71

N= 56

148 A

Typisch für die traumatische intrakranielle Blutung sind: (Wähle zwei)

- 1- zunehmende Bewußtseinstörung
- 2- gleichbleibende Bewußtseinstörung
- 3- langsame Aufhellung des Bewußtseins
- 4- nach Aufhellung des Bewußtseins wieder eintretende Bewußtseinstörung (freies Intervall)
- 5- volles Bewußtsein

Testform B

M= 1.58

N= 59

105 B

Welche der genannten Symptome sind typisch für eine Colitis ulcerosa? (Wähle zwei)

- 1- chronische, milde Diarrhoen
- 2- Steatorrhoe mit reichlich unverdaulichem Eiweiß und Stärke
- 3- unstillbare Diarrhoen mit viel unverdaulichem Nahrungsresten
- 4- unter Tenesmen entleerte, blutige-schleimige Stühle, oft ohne fäkulente Beimengungen
- 5- Melaena
- 6- Dehydratation
- 7- pathologischer Sekretintest

Für den Subtest "Diagnose mit technischen Hilfsmitteln" hingegen ergab sich keine einheitliche Bewertungstendenz für beide Testformen. So scheinen in Test A die Fragen eher negativ, in Test B dagegen, wenn überhaupt vom Durchschnitt abweichend, so eher positiv bewertet worden sein. Sucht man nach einer Erklärung für diese allerdings nicht ausreichend signifikant belegte schlechtere Bewertung des Test A, so fallen mehrere Items auf, die relativ komplizierte biochemische Fragestellungen enthalten, wie z.B. die Fragen 6 A oder 125 A.

Fragenbeispiele aus dem Subtest "Diagnose mit technischen Hilfsmitteln"

Testform A M= 2.04
 N= 56

6 A

Welche der folgenden Kombinationen von pathologischen Veränderungen charakterisiert die respiratorisch bedingte Alkalose? (Wähle eine)

- 1- Blut pH und Urin pH höher als 7,4;
niedriger CO_2 -Gehalt des Blutes; erhöhte Natriumkonzentration im Urin; geringe CO_2 -Bindungskapazität.
- 2- Blut pH geringer, Urin pH höher als 7,4;
niedriger CO_2 -Gehalt des Blutes; erhöhte Natriumkonzentration im Urin; geringe CO_2 -Bindungskapazität.
- 3- Blut pH höher, Urin pH geringer als 7,4;
niedriger CO_2 -Gehalt des Blutes; hohe Natriumkonzentration im Urin; geringe CO_2 -Bindungskapazität.
- 4- Blut und Urin pH höher als 7,4; hoher Blut CO_2 -Gehalt; zunehmende Natriumkonzentration im Urin; geringe CO_2 -Bindungskapazität des Blutes.
- 5- Blut pH und Urin pH höher als 7,4; niedriger CO_2 -Gehalt des Blutes; abnehmende Natriumkonzentration im Urin; geringe CO_2 -Bindungskapazität des Blutes.

125 A

Erhöhte Kalziumausscheidung im Urin kommt vor bei: (Wähle zwei)

- 1- Parathyreopriver Tetanie
- 2- seniler Osteoporose
- 3- Osteogenesis imperfecta tarda
- 4- Hyperparathyreoidismus
- 5- Osteomalazie
- 6- "Boeck'scher" Sarkoidose

Eine Kategorie, für die sich wieder übereinstimmend in beiden Testformen eine Tendenz zu positiver Bewertung abzeichnet, stellt der Subtest "Prognose" dar, der durch die folgenden Fragenbeispiele charakterisiert werden kann. Frage 51 A taucht dabei erneut auf, die bereits wegen ihrer positiven Bewertung innerhalb der ätiologischen Kategorie der Infektionen erwähnt wurde.

Fragenbeispiel aus dem Subtest "Prognose"

Testform A

M= 1.63
N= 16

51 A

- 1- Erysipel
- 2- Lobärpneumonie
- 3- Meningitis
- 4- Scharlach
- 5- abszedierende Entzündung
- 6- Endocarditis lenta
- 7- hämorrhagische Pneumonie
- 8- Waterhouse-Friderichsen-Syndrom

Welche der oben angeführten Krankheiten ist in nahezu allen Fällen ihres Auftretens die Komplikation einer Grippeerkrankung? (Wähle eine)

127 A

Welche der aufgeführten Erkrankungen sind typische Komplikationen des unbehandelten Scharlachs? (Wähle zwei)

- 1- Glomerulonephritis
- 2- Bronchopneumonie
- 3- eitrige Otitis media
- 4- Pankreatitis
- 5- Enzephalitis

Testform B

M= 1.45
N= 20

62 B

Die Prognose einer operierten Magenperforation hängt in erster Linie ab von: (Wähle eine)

- 1- Anaesthesieverfahren
- 2- Lokalisation der Perforation
- 3- Zeit bis zur Operation
- 4- Wahl des Operationsverfahrens (Übernähung; Magenresektion)
- 5- der sorgfältigen Operationsvorbereitung

172 B

- 1- vorzeitige Placentalösung
- 2- Randsinus-Blutung
- 3- Placenta praevia
- 4- Uterusruptur
- 5- Zeichnen

Welche der obigen Erscheinungen droht bei Hochsteigen der Bandlischen Furche, Druckschmerzhaftigkeit zwischen Nabel und Symphyse und starker, pausenloser Wehentätigkeit? (Wähle eine)

Für beide Tests liegt jedoch die Zufallswahrscheinlichkeit der Verteilung der Fragen dieser Kategorie auf die Mittelwertsklassen oberhalb 5%.

Ähnliches gilt für die Fragen des Subtests "Gastrologie-Enterologie", wo jedoch wiederum allein die übereinstimmende Tendenz für beide Testformen, nicht jedoch das Signifikanzniveau der einzelnen Verteilungen die Erwähnung des Ergebnisses sinnvoll erscheinen läßt.

Fragenbeispiele aus dem Subtest "Gastrologie-Enterologie"

Testform A

M= 1.66
N= 15

64 A

Ein 64 Jahre alter Mann klagt über Dysphagie, Sodbrennen und 4 kg Gewichtsverlust. Röntgenaufnahmen ergeben eine mäßige, unregelmäßige Verengung des distalen Oesophagus. Die wahrscheinliche Ursache ist:
(Wähle eine)

- 1- Oesophagitis oder Karzinom
- 2- Oesophagusvarizen oder Gefäßläsionen
- 3- Mallory-Weiss-Syndrom oder Karzinom
- 4- Kardiospasmus oder funktionale Obstruktion
- 5- kongenitale Striktur oder Oesophagitis
- 6- gutartige Geschwulst oder Kardiospasmus

Testform B

M= 1.65
N= 17

34 B

Ein 9 Monate altes Kind wird nachts in die Klinik gebracht, weil es weinend aufwachte, dabei seine Beine hochzog und anscheinend kolikartige Bauchschmerzen hatte. Es hat kein Fieber und keinen Durchfall, aber einen blutigen Stuhl. Welche der folgenden ist die naheliegende Verdachtsdiagnose? (Wähle eine)

- 1- Infektion mit Shigellen
- 2- blutendes Ulcus pepticum
- 3- Invagination
- 4- Meckelsches Divertikel
- 5- Divertikulitis

Damit sind im wesentlichen diejenigen Subtests dargestellt, für die sich mehr oder minder deutlich eine Tendenz zu positiven Bewertungen abzeichnete.

6.2.4.2 Negativ bewertete Fragenkategorien

Der einzige Befund, der für beide Testformen deutlich signifikant zu erheben war, tauchte dagegen bei den Fragen auf, die eher negativ bewertet wurden. Und zwar waren es hier die Fragen ohne ärztlichen Anwendungsbereich, die in beiden Tests besonders schlecht bewertet wurden. Es handelt sich bei diesen Aufgaben u. a. um theoretische Fragestellungen aus dem Bereich der Genetik, Biochemie oder Pathologie.

Fragenbeispiele aus dem Subtets "ohne ärztliche Anwendung"

Testform A

M= 2.73
N= 22

36 A

- 1- XO; 45 Chromosomen
- 2- XY; 46 Chromosomen
- 3- XX; 46 Chromosomen
- 4- Trisomie-21; 47 Chromosomen
- 5- XXY; 47 Chromosomen
- 6- XY; 48 Chromosomen
- 7- Trisomie -17 und -18;
47 Chromosomen
- 8- 49 Chromosomen

Welche der oben angeführten Chromosomenkonstellationen findet sich bei einem 12 Jahre alten Knaben mit hypophysärem Minderwuchs?
(Wähle eine)

147 A

Welche der folgenden Hirngeschwülste treten vorwiegend im Kindes- und Jugendalter auf?
(Wähle zwei)

- 1- Glioblastoma multiforme
- 2- Medulloblastom
- 3- Ependymom
- 4- Hypophysenadenom
- 5- Oligodendrogliom
- 6- malignere Form des Astrocytoms

Testform B

M= 2.63
N= 16

38 B

Welches Ferment ist für die Eliminierung des Bilirubins beim Neugeborenen erforderlich?
(Wähle eine)

- 1- Laktat-Dehydrogenase
- 2- alkalische Phosphate
- 3- Glukuronyl-Transferase
- 4- Serumglutamatpyruvat-Transaminase
- 5- Glukose-6-Phosphat-Dehydrogenase

Die Zufallswahrscheinlichkeit der Mittelwertverteilung für diese Kategorie liegt für beide Testformen unter 0,1%.

Eine weitere Kategorie für die sich ganz deutlich wenigstens in Test B - eine Tendenz zu negativer Bewertung abzeichnet, sind die Fragen des Subtests "Kongenitale Anomalie". Diese Kategorie enthält Fragen über angeborene Fehlbildungen des Herzens des Skeletts und anderer Organsysteme. In Test B, der besonders schlecht abschneidet, fällt eine Häufung von Aufgaben über mehr theoretische, erbbiologische Themen auf.

Fragenbeispiele aus dem Subtest "Kongenitale Anomalie"

Testform A

M= 2.21
N = 14

44 A

Zu welchem der folgenden Herzvitien gehört die Linkshypertrophie NICHT? (Wähle eine)

- 1- Ventrikelseptumdefekt
- 2- offener Ductus Botalli
- 3- Trikuspidalatresie
- 4- Vorhofseptumdefekt
- 5- Aortenstenose

151 A

Ein 18 Monate altes Mädchen wird zur Begutachtung in die Sprechstunde des Hausarztes gebracht. Seitdem das Kind vor etwa einem Monat zu laufen begann, wurde ein linksseitiges, schmerzloses Hinken beobachtet. Das Kind war voll ausge tragen worden, es handelte sich jedoch um eine Steißlage. Zwei Geschwister sind normal. Eine 33 Jahre alte Tante mütterlicherseits hat einen kontinuierlich zunehmenden Schmerz rechts seit etwa 4 Jahren. Sonst ist das Kind seit Geburt in ausgezeichnetem Gesundheitszustand.

Welche der folgenden Befunde könnte man bei der klinischen Untersuchung erwarten? (Wähle 2)

- 1- eingeschränkte Abduktion der linken Hüfte, symmetrische Glutealfalten
- 2- der Trochanter steht oberhalb der Roser-Nélatonschen Linie (Verbindungslinie zwischen Spina iliaca anterior superior und Tuber ossis ischii)

Testform B

M= 2.91
N = 11

39 B

Folgende Erkrankungen werden auf einen Enzymdefekt zurückgeführt, außer: (Wähle eine)

- 1- Fruktoseintoleranz
- 2- adrenogenitales Syndrom
- 3- Galaktosämie
- 4- Fölling'sche Krankheit (Phenylketonurie)
- 5- Morbus Hand-Schüller-Christian

88 B

- 1- Turner-Syndrom
- 2- Testikuläre Feminisierung
- 3- Klinefelter-Syndrom
- 4- adrenogenitales Syndrom
- 5- keines der genannten

Zu welchem der genannten Krankheitsbilder gehören folgende Symptome:

positives Sexchromatin, chromosomales Geschlecht XX, normale Östrogenausscheidung, hohe 17-Ketosteroidausscheidung, normale innere Genitalien? (Wähle eine)

124 B

Welche sind die ersten diagnostischen Maßnahmen bei einem neugeborenen, das beständig bläulichen Speichel vor dem Munde hat, weil es offenbar seinen Speichel nicht schlucken kann? (Wähle zwei)

- 1- Röntgen-Leeraufnahme im Hängen
- 2- Sondierungsversuch

- | | |
|--|---|
| 3- eingeschränkte Streckung
der linken Hüfte, Abduktion
normal, asymmetrische
Glutealfalten
4- eingeschränkte Abduktion
und Streckung der linken Hüfte
mit asymmetrischen Gluteal-
falten
5- die proximale Spitze des linken
großen Trochanter steht distal
der Roser-Nélatonschen Linie | 3- Kontrastmitteldarstellung
4- Ösophaguskopie
5- Fütterungsversuch |
|--|---|

Ähnlich wie bei der zuvor genannten Kategorie verhält sich auch bei den Fragen aus den Querkategorien Blutkrankheiten und Endokrinologie-Metabolismus. Es ergibt sich jeweils eine signifikante negative Bewertungstendenz in Test B bei unsignifikanter, jedoch tendenziell gleichsinniger Bewertung in Test A. In der Kategorie Blutkrankheiten findet sich eine Reihe ausgesprochen theoretischer Fragen, wie z.B. die Aufgaben 8 B oder 10 B, die schlechte Bewertungen erhielten, während in Test A die Fragen immerhin doch praxisnäher erscheinen, wie die Aufgaben 42 A oder 115 A zeigen.

Fragenbeispiele aus dem Subtest "Blutkrankheiten"

Testform A

M= 2.00
N= 7

42 A

Welche der folgenden ist die häufigste Ursache der Anämie bei Kleinkindern? (Wähle eine)

- 1- Vitaminmangel
- 2- Eingeweidewürmer
- 3- rezidivierendes Erbrechen
- 4- Dystrophien
- 5- Eisenmangel

115 A

Ein 55-jähriger Mann klagt über Kopfschmerzen, Schwindel, Müdigkeit und Schwäche. Er bemerkt auch eine zunehmende Rötung des Gesichts. Sonst hat er keine Beschwerden. Befund: Das Gesicht erscheint etwas gerötet; die Finger und Zehen sind leicht zyanotisch. Die Leber ist eben, die Milz bei tiefer Inspira-

Testform B

M= 2.56
N= 9

8 B

Bei der Granulopoese lassen sich folgende Zellformen unterscheiden:

- A Metamyelocyt
- B Myelocyt
- C Myeloblast
- D Promyelocyt

Welche der folgenden Anordnungen gibt die chronologische Reihenfolge des Entstehens dieser Zellen richtig wieder? (Wähle eine)

- 1- C A B D
- 2- A C D B
- 3- D C B A
- 4- C D A B
- 5- C D B A
- 6- A B C D

tion deutlich tastbar. Krankhafte Befunde an Herz und Lunge bestehen nicht.

Blutbild: Hb 20 g%, Erythrozyten 7,5 Mill., Leukozyten 16 500, Thrombozyten 800 000, Hämatokrit 68%, Differential-Blutbild unauffällig. BSG 1/2mm.

Die wahrscheinliche Diagnose ist: (Wähle eine)

- 1- Polycythaemia vera
- 2- Dünndarm-Karzinoid mit Flush-Syndrom
- 3- symptomatische Polyglobulie
- 4- chronisch-myeloische Leukämie
- 5- Osteomyelosklerose

10 B

Sie bekommen ein Teströhrchen mit frisch abgezogenem Blut und werden gebeten, es zu beobachten. Sie bemerken, daß sich ganz normal ein Gerinnsel bildet und, daß sich dieses innerhalb einer halben Stunde retrahiert. Nach drei Stunden Aufbewahren bei 37°C aber zeigt sich, daß sich das Koagulum vollkommen aufgelöst hat.

Welche Befunde würden Sie bei diesem Patienten erwarten? (Wähle eine)

- 1- Rückenschmerzen, Ekchymosen und eine Orchiektomie
- 2- Polyurie und eine Amputation oberhalb des Knies
- 3- starke Halsvenenstauung, leichte Atemnot und Hypertonie
- 4- Blässe, diffuse Petechien, Splenomegalie und ein pathologischer neurologischer Befund. Im peripheren Blut-Ausstrich keine Plättchen

Bei den Fragen der Kategorie Endokrinologie-Metabolismus handelt es sich im Test B fast ausschließlich um praxisferne Items, wie z.B. die Aufgabe 39, 137 oder 71 B. Nur eine Frage behandelt die Hyperthyreose, keine einzige den Diabetes mellitus. In Test A dagegen behandeln etwas weniger als die Hälfte der Fragen Pankreas bzw. Schilddrüsenstoffwechsel, wie z.B. die Aufgaben 131 oder 122 A, so daß dieser Subtest relativ besser abschneidet.

Fragenbeispiele aus dem Subtest "Endokrinologie-Metabolismus"

Testform AM= 2.09
N= 22122 A

Durch welche der folgenden Symptome stellt sich eine Hypothyreose bei einer Frau in der Menopause dar? (Wähle zwei)

- 1- psychische Teilnahmslosigkeit
- 2- Abmagerung
- 3- Bradykardie
- 4- erhöhte Fixierung von Jod 131

131 A

Dringliche Sofortmaßnahmen bei diabetischem Koma sind: (Wähle drei)

- 1- intravenöse Traubenzuckerinfusion
- 2- Salz- und Wasserersatz
- 3- Insulininjektion
- 4- Acidosebehandlung
- 5- Infektionsprophylaxe

Testform BM= 2.41
N= 1739 B

Folgende Erkrankungen werden auf einen Enzymdefekt zurückgeführt, außer: (Wähle eine)

- 1- Fruktoseintoleranz
- 2- adrenogenitales Syndrom
- 3- Galaktosämie
- 4- Fölling'sche Krankheit (Phenylketonurie)
- 5- Morbus Hand-Schüller-Christian

71 B

Der traumatische Diabetes insipidus kann hervorgerufen werden durch direkte oder indirekte Verletzung von: (Wähle eine)

- 1- Pankreas
- 2- beiden Nieren
- 3- Hypothalamus
- 4- Frontalhirn
- 5- Nebennieren

137 B

- 1- eisenfrei
- 2- eisenhaltig
- 3- intrazellulär gelegen
- 4- extrazellulär gelegen
- 5- identisch mit Bilirubin
- 6- mit Ferritin gekoppelt

Bei Zerfall von roten Blutkörperchen kann Hämosiderin entstehen. Welche der oben angeführten Eigenschaften kommen diesem Pigment zu? (Wähle drei)

Für beide Tests in der Tendenz praktisch gleich, jedoch unterhalb der 5% Signifikanzgrenze, ist die Verteilung der Bewertung in der Kategorie "OP.-Therapie und Anästhesie", deren Fragen von den Ärzten eher negativ bewertet wurden. Als Beispiele für diese Kategorie seien die Aufgaben 179 A sowie 155 und 62 B angeführt.

Fragenbeispiele aus dem Subtest "Operative Therapie-Anästhesie"

Testform A

M= 2.27
N= 11

179 A

Welche der folgenden Maßnahmen sind geeignet bei einem invasiven Collum-Karzinom der Gruppe I B?
(Wähle zwei)

- 1- vaginale Uterusexstirpation
- 2- abdominale Uterusexstirpation
- 3- Wertheimsche Radikalooperation mit Entfernung der Beckenlymphknoten
- 4- Radium- und Röntgenbestrahlung
- 5- keine der obigen

Testform B

M= 2.25
N = 12

62 B

Die Prognose einer operierten Magenperforation hängt in erster Linie ab von: (Wähle eine)

- 1- Anaesthesieverfahren
- 2- Lokalisation der Perforation
- 3- Zeit bis zur Operation
- 4- Wahl des Operationsverfahrens (Übernähung; Magenresektion)
- 5- der sorgfältigen Operationsvorbereitung

155 B

- 1- Verminderung des Herzminutenvolumens
- 2- postoperative Azidose
- 3- Schleimhautreizung, vermehrte Bronchialsekretion
- 4- Exzitation
- 5- zu geringe Analgesie

Welche der angeführten Wirkungen sind besonders bei der Anwendung von Äther zu Narkosezwecken zu erwarten? (Wähle drei)

In der Kategorie "physikalische Einwirkungen" schneiden die Fragen aus Test A etwas schlechter ab als in Test B. Hier betreffen allein zwei der sechs Fragen Probleme der postoperativen Behandlung (69 A) und auch im Bereich der Traumatologie scheinen die Fragen des Test B der Primärsituation gemäß (57 B) als die des Test A (31 A).

Fragenbeispiele aus dem Subtest "physikalische Einwirkungen"

Testform A

M= 2.17
N= 6

31 A

Bei einem Säugling besteht der Verdacht auf ein subdurales Hämatom bzw. einen subduralen Erguß. Welche der folgenden diagnostischen Maßnahmen würden Sie als ERSTE ergreifen? (Wähle eine)

- 1- Luftenzephalogramm
- 2- Lumbalpunktion
- 3- Ventrikelpunktion
- 4- Ventrikulographie
- 5- Fontanellenpunktion

69 A

Eine 47 Jahre alte Frau wird einer abdominoperinealen Resektion unterzogen. Während der Operation ist die Patientin für 40 Min. hypotonisch. In den ersten sechs Stunden nach der Operation ist die Urinausscheidung vermindert, im Durchschnitt 5 ml/Std. Der Blutdruck ist durchschnittlich 130/70 mmHg (präoperativer Blutdruck: 140/80 mmHg) die Pulsfrequenz ist 100/min.

Die Ursache für die Oligurie ist höchstwahrscheinlich: (Wähle eine)

- 1- postoperativer Schock
- 2- akutes Nebennierenversagen
- 3- akute tubuläre Nekrose
- 4- Unterbindung eines Ureters
- 5- Unterbindung beider Ureter

Testform B

M= 1.91
N= 11

57 B

Der Patient ist ein 55-jähriger Krankenpfleger. Um 7.30 Uhr stürzte er von seinem Fahrrad auf den Kopf. Er war benommen, aber nicht bewußtlos. Er ging nach Hause. Dort stellte sich gegen 11.00 Uhr Somnolenz und Bewußtlosigkeit ein. Er wurde in die Klinik gebracht. Das Röntgenbild zeigt: Fissur links parietal. Anisokorie: Linke Pupille größer als rechte. Welche ist die wahrscheinlichste Diagnose? (Wähle eine)

- 1- Comotio cerebri mit Schädel-fissur
- 2- Verdacht auf subdurales Hämatom
- 3- Verdacht auf epidurales Hämatom
- 4- Verdacht auf Stammhirnblutung
- 5- posttraumatischer Kollaps bei Arteriosklerose und Schädeltrauma

Als Erklärung für das bessere Abschneiden des Test B in der Kategorie "Allergene, Antigene" fällt hier eine Anzahl positiv bewerteter Fragen auf, die allergische Phänomene bei Stryptokokkeninfektionen und deren Therapie behandeln (126, 30 B) gegenüber theoretischen Fragen wie dem Items 21 in Test A.

Fragenbeispiele aus dem Subtest "Allergene, Antigene"

Testform A

M= 2.25
N= 4

21 A

Bei den allergischen Sofortreaktionen (z.B. dem anaphylaktischen Schock bei Meeresschweinchen, der akuten Urtikaria, dem Heuschnupfen und dem jahreszeitlich auftretenden Bronchialasthma) treten die Symptome auf, wenn der Antigen-Antikörperkomplex chemische Substanzen freisetzt. Während der Mastzellendegranulation z.B. setzen die Granula - die Histamin mit Heparin enthalten - Histamin frei, wodurch die glatte Muskulatur und die Speichelsekretion stimuliert werden und die Kapillarpermeabilität gesteigert wird. Die therapeutischen Bemühungen sind darauf gerichtet, das Histamin entweder auf der Rezeptorseite des allergisch reagierenden Gewebes zu blockieren oder den Histamineffekt aufzuheben. In letzter Zeit wurden von mehreren pharmazeutischen Firmen Präparate in den Handel gebracht, die versuchen, die Produktion von Histamin durch die Mastzellen zu verhindern. Um dies zu erreichen, braucht man einen einigermaßen spezifischen: (Wähle eine)

- 1- Sulphydrylinhibitor
- 2- Decarboxylaseinhibitor
- 3- Folsäureantagonisten
- 4- Hyaluronidaseinhibitor
- 5- eine Antiribonuclease

Testform B

M= 1.88
N= 8

126 B

Ein 7 Jahre altes Mädchen hat eine nachweisbar durch β -hämolytische Streptokokken bedingte Tonsillitis. Nach Angaben der Mutter habe eine frühere Gabe von Penicillin zu einer Urticaria geführt. Welche der folgenden Medikamente darf man dem Kind NICHT geben? (Wähle zwei)

- 1- Tetracycline für 10 Tage
- 2- Natriumoxacillin (z.B. Stapenor) für 10 Tage
- 3- Chloramphenicol bis der bakteriologische Befund negativ ist
- 4- Erythromycin für 10 Tage
- 5- Sulfisoxazol (Gantrisin) für 10 Tage
- 6- Penicillin und Benadryl (ein Phenothiazinderivat) oral für 10 Tage

30 B

Anmerkung: Die in der Tabelle dargestellten Urinbefunde 1 bis 6 sind Alternativen zu den folgenden 3 Aufgaben.

	pH	Protein	Zucker	Aceton	Sediment
1-	6.0	(+)	0	0	20 bis 30 Leukozyten pro Gesichtsfeld, zum Teil agglomeriert
2-	6.0	4+	0	0	5 bis 10 hyaline und 3 bis 6 granulierte Zylinder, doppelbrechende Substanzen
3-	6.0	3+	0	0	viele Erythrozyten, 5 bis 10 Leukozyten, 2 bis 4 granulierte und 0 bis 2 Erythrozytenzylinder
4-	5.0	2+	4+	4+	3 bis 5 Leukozyten pro Gesichtsfeld
5-	6.5	(+)	0	4+	Schleimfäden, einige Oxalokristalle
6-	7.5	0	0	0	massenhaft Bakterien, einige Oxalatkristalle

Welche der sechs Urinproben könnte der Morgenurin eines 4-jährigen Mädchens mit einer Hypertonie, periorbitalen und Knöchelödemen und mit einem erhöhten Antistreptolysinintiter sein? (Wähle eine)

6.2.4.3 Fragenkategorien, für die sich in den beiden Testformen entgegengesetzte Bewertungstendenzen abzeichnen

Neben den zuvor dargestellten Fragenkategorien, für die sich mehr oder minder deutliche und für beide Testformen gleichsinne Bewertungstendenzen finden ließen, fallen schließlich noch solche auf, bei denen einer eher positiveren Bewertung der Fragen des einen Tests eine eher negative derjenigen des anderen gegenübersteht. So finden sich in der Querkategorie Niere, Urologie deutliche Unterschiede in der Bewertung der Fragen aus Test A bzw. B. Der Vergleich zeigt als wahrscheinliche Ursache für die positive Bewertung der B-Fragen die Häufung von Fragen nach einfachen aber wesentlichen Sachverhalten im Zusammenhang mit Nierenerkrankungen (114 B, 13 B). Für die schlechtere Bewertung des Tests A dagegen sind 2 Fragen aus der speziellen postoperativen Nephrologie (69, 70 A) neben theoretischeren Fragen wie Aufgabe 166 A mitverantwortlich.

Fragenbeispiele aus dem Subtests "Niere-Urologie"

<u>Testform A</u>	M= 1.94 N = 16	<u>Testform B</u>	M= 1.31 N = 13
<u>69 A</u>		<u>13 B</u>	
Eine 47 Jahre alte Frau wird einer abdominoperinealen Resektion unterzogen. Während der Operation ist die Patientin für 40 Min. hypotonisch. In den ersten sechs Stunden nach der Operation ist die Urinausscheidung vermindert, im Durchschnitt 5 ml/Std. Der Blutdruck ist durchschnittlich 130/70 mmHg (präoperativer Blutdruck: 140/80 mmHg), die Pulsfrequenz ist 100/min.		Nach welcher der folgenden Krankheiten tritt häufig eine Glomerulonephritis auf? (Wähle eine)	
		1- Mumps	
		2- Diphtherie	
		3- Pocken	
		4- Scharlach	
		5- Masern	
Siw Ursache für die Oligurie ist höchstwahrscheinlich: (Wähle eine)		<u>114 B</u>	
1- postoperativer Schock		Welche der folgenden Symptome sind CHARAKTERISTISCH für die akute, diffuse Glomerulonephritis? (Symptome, die in der Regel auch bei der parainfektiosen Herdnephritis vorkommen, NICHT markieren.) (Wähle zwei)	
2- akutes Nebennierenversagen		1- Proteinurie	
3- akute tubuläre Nekrose		2- Hämaturie	
4- Unterbindung eines Ureters		3- Hypertonie	
5- Unterbindung beider Ureter		4- Ödeme	
<u>70 A</u>			
Welche der folgenden Maßnahmen könnte zum Nachweis der Ursache dieser postoperativen Oligurie			

nützlich sein? (Wähle eine)

- 1- Gabe von Noradrenalin
- 2- Gabe von Hydrocortison
- 3- Gabe von Phenolsulfonphtalein
- 4- Gabe von Mannit
- 5- Messung des zentralen Venendruckes

166 A

Welche Hormone steuern die tubuläre Rückresorption von Wasser oder Natrium? (Wähle zwei)

- 1- Vasopressin (ADH=antidiuretisches Hormon)
- 2- Aldosteron
- 3- Insulin
- 4- Oxytocin
- 5- Parathormon

Auch bei den Fragen aus der Querkategorie Respirationssystem fällt für Test A eine negative Bewertungstendenz auf im Gegensatz zu Test B, der eine eher positive Bewertung erfuhr. Die schlecht bewerteten Fragen aus Test A, die für dieses Ergebnis hauptverantwortlich sind, sind die Fragen 6 und 68. Beispiele sehr positiver Fragen aus Test B sind die Aufgaben 68 und 108.

Fragenbeispiele aus den Subtests "Respirationssystem"

Testform A

M= 2.00
N= 10

Testform B

M= 1.71
N= 7

6 A

Welche der folgenden Kombinationen von pathologischen Veränderungen charakterisiert die respiratorisch bedingte Alkalose? (Wähle eine)

- 1- Blut pH und Urin pH höher als 7,4; niedriger CO₂-Gehalt des Blutes; erhöhte Natriumkonzentration im Urin; geringe CO₂-Bindungskapazität.
- 2- Blut pH geringer, Urin pH höher als 7,4; niedriger CO₂-Gehalt des Blutes; erhöhte Natriumkonzentration im Urin; geringe CO₂-Bindungskapazität.
- 3- Blut pH höher, Urin pH geringer als 7,4; niedriger CO₂-Gehalt des Blutes; hohe Natriumkonzentration im

68 B

Bei Verdacht auf Bronchialkarzinom ist zur Erhärtung der Diagnose NICHT zu empfehlen: (Wähle eine)

- 1- Röntgen in 2 Ebenen, Tomographie
- 2- Bronchoskopie
- 3- cytologische Untersuchung
- 4- Probethorakotomie
- 5- Kontrolle des Befundes in kurzen Zeitabständen

Urin; geringe CO_2 -Bindungs-
kapazität.

- 4- Blut und Urin pH höher als 7,4;
hoher Blut CO_2 -Gehalt; zuneh-
mende Natriumkonzentration im
Urin; geringe CO_2 -Bindungska-
pazität des Blutes.
- 5- Blut pH und Urin pH höher als 7,4;
niedriger CO_2 -Gehalt des Blutes;
abnehmende Natriumkonzentration
im Urin; geringe CO_2 -Bindungs-
kapazität des Blutes.

68 A

Ein neun Stunden altes Neugeborenes ist zyanotisch und in Atemnot. Die Atemfrequenz ist 96/min. Die Atemgeräusche über der linken Thorax-
hälfte sind gedämpft, der Herzspitzenstoß wird in der Mittellinie be-
obachtet.

Die wahrscheinliche Diagnose ist:
(Wähle eine)

- 1- Vorhofseptumdefekt
- 2- Zwerchfellhernie
- 3- offener Ductus arteriosus
- 4- Oesophagotrachealfistel

In der Querkategorie Cardiologie-Angiologie schnitten die Fragen des Tests B etwas schlechter ab. Besonders schlecht bewertet wurden hier die Fragen 27, 33 und 50. Demgegenüber fällt in Test A eine stärkere Häufung von Fragen über herz- und blutdruckwirksame Pharmaka auf, die sehr positiv bewertet wurden, wie z.B. Aufgabe 79 oder 77.

Fragenbeispiele aus den Subtests "Cardiologie-Angiologie"

Testform A

M= 1.90
N = 21

77 A

Bei welchem der folgenden Herz-
glykoside ist die kumulierende Wir-
kung am stärksten ausgeprägt?
(Wähle eine)

- 1- Digitoxin
- 2- Digoxin
- 3- Lanatosid C
- 4- Strophanthin
- 5- Digilanid
- 6- Scillaren

108 B

Als typisch für das Sputum eines
erwachsenen Asthmaticus gel-
ten: (Wähle zwei)

- 1- serös-flüssige Konsistenz
- 2- Vorhandensein elastischer
Fasern
- 3- Vorhandensein jugendlicher
Leukozyten
- 5- schleimig-zähe Konsistenz

Testform B

M= 1.90
N = 21

27 B

Welche der folgenden Gehirn-
veränderungen ist mit einem
Naevus flammeus des Gesichtes
kombiniert und verrät sich fast
immer durch Krämpfe?
(Wähle eine)

- 1- Kraniopharyngeom
- 2- Morbus Sturge-Weber
- 3- Astrocytom
- 4- Medulloblastom
- 5- Ependymom

79 A

Welche der angeführten Gruppen von Diuretica ist außer zur Ödemausschwemmung zur Blutdrucksenkung bei Hypertonikern geeignet?
(Wähle eine)

- 1- Purinkörper
- 2- Quecksilberdiuretica
- 3- Aldosteronantagonisten
- 4- Chlorothiazidderivate
- 5- Carbonanhydrasehemmstoffe
- 6- osmotisch wirkende Diuretica

33 B

Womit ist eine im Phonokardiogramm darstellbare, weite und fixierte Spaltung des zweiten Herztones häufig kombiniert?
(Wähle eine)

- 1- Vorhofseptumdefekt
- 2- Ventrikelseptumdefekt
- 3- offener Ductus Botalli
- 4- Fallotsche Tetralogie
- 5- Transposition der großen Gefäße

50 B

Bei welchen der folgenden Symptome ist eine Durchtrennung eines offenen Ductus arteriosus KON-
TRAINDIZIERT? (Wähle eine)

- 1- keine Symptome
- 2- Herzversagen
- 3- Symptome einer subakuten und therapeutisch nicht beeinflussbaren bakteriellen Endocarditis
- 4- Zyanose und Rechts-Links-Shunt
- 5- Links-Rechts-Shunt

Eine deutlich schlechtere Bewertung der Fragen des Test B findet sich auch in der Querkategorie Nervensystem. Zur negativeren Bewertung tragen hier hauptsächlich einige Fragen aus der Anästhesiologie (155 B) sowie aus der topographischen oder funktionellen Neuroanatomie (46 B) bei. Gut bewertet wurden dagegen Fragen über psychotrope Pharmaka (157 A, 77 B), die im Test A sehr viel stärker vertreten sind, ebenso wie Fragen, die diagnostisch wichtige neurologische Symptome behandeln (101 A, 120 B).

Fragenbeispiele aus dem Subtest "Nervensystem"

Testform A

M= 1.84

N= 19

101 A

- A Bei der akuten Poliomyelitis bestehen KEINE Störungen der Sensibilität,
weil
- B die Pyramidenbahnen von dieser Erkrankung NICHT ergriffen werden

157 A

Welche der folgenden Behauptungen treffen zu? (Wähle zwei)

- 1- Phenobarbital (Luminal, Wz.) hemmt Grand-mal-Anfälle
- 2- Phenobarbital hemmt Petit-mal-Anfälle
- 3- Phenobarbital wirkt in nicht-sedativen Dosen analgetisch
- 4- Phenobarbital dringt NICHT durch die Plazentarschranke
- 5- Phenobarbital verursacht gelegentlich Erregungszustände

Testform B

M= 2.11

N= 18

46 B

Eine Läsion des Nervus axillaris unmittelbar nach dem Abgang aus dem Fasciculus dorsalis führt zu einer Lähmung des: (Wähle eine)

- 1- Musculus coracobrachialis
- 2- Musculus brachialis
- 3- Musculus triceps brachii
- 4- Musculus deltoideus
- 5- Musculus biceps brachii

77 B

Welches Arzneimittel eignet sich zur medikamentösen Unterstützung der Therapie agitierter Psychosen? (Wähle eine)

- 1- Morphin
- 2- Chlorpromazin (Megaphen, Wz.)
- 3- Diphenylhydantoin (Zentropil, Wz.)
- 4- Imipramin (Tofranil, Wz.)
- 5- Monoaminoxidasehemmstoffe, (z.B. Niamid, Wz.)

120 B

Welche der genannten Erscheinungen finden Sie beim Morbus Parkinson? (Wähle drei)

- 1- erhöhter Muskeltonus
- 2- abnorme Bewegungen
- 3- mimische Starre
- 4- Empfindungsdissoziation

155 B

Welche der angeführten Wirkungen sind besonders bei der Anwendung von Äther zu Narkosezwecken zu erwarten? (Wähle drei)

- 1- Verminderung des Herzminutenvolumens
- 2- postoperative Azidose
- 3- Schleimhautreizung, vermehrte Bronchialsekretion
- 4- Exitation
- 5- zu geringe Analgesie

Bei der ätiologischen Kategorie "Neoplastischer Prozeß" findet sich eine negative Bewertung der Fragen aus Test A bei schwach positiver Tendenz in Test B. Als Erklärung hierfür fällt zum einen auf, daß der Test B mehr praktisch diagnostische Fragen beispielsweise zum weiblichen Genitalkarzinom enthält (173, 175), die positiv bewertet wurden, während der Test A mehr operative therapeutische Aspekte in seinen Fragen enthält (180). Zum anderen betrifft ein Teil der Fragen des Test A seltenere Erkrankungen oder rein theoretische Fragestellungen, wie z. B. die Aufgabe 113 oder 50.

Fragebeispiele aus den Subtests "Neoplastischer Prozeß"

<u>Testform A</u>	M = 2.09 N = 22	<u>Testform B</u>	M = 1.86 N = 21
-------------------	--------------------	-------------------	--------------------

113 A

Welche der folgenden Veränderungen gehören zum Zollinger-Ellison Syndrom? (Wähle zwei)

- 1 - Vermehrung der insulinbildenden Beta-Zellen im Pankreas
- 2 - starke Vermehrung der Magensaftsekretion
- 3 - Magen-Oesophagus-Duodenal-ulcera
- 4 - Polyposis des Colons mit Osteomen, Fibromen und Talgzysten
- 5 - generalisierte intestinale Polyposis mit Melaninflecken in der Wangenschleimhaut, an den Lippen und an den Fingern

50 A

Die Energiegewinnung in Zelle artiger Geschwülste beruht in erster Linie auf: (Wähle eine)

- 1 - aerober Glykolyse
- 2 - anaerober Glykolyse
- 3 - Verwertung von Stickstoff
- 4 - Verwertung von Kohlendioxyd
- 5 - Verwertung von Fett

180 A

Welche der folgenden Behandlungsmethoden kommen infrage bei einem nach Konisation festgestellten frühinvasiven Karzinom der Cervix bei einer 34jährigen verheirateten Frau mit 2 Kindern? (Wähle drei)

173 B

Welche der folgenden Feststellungen stimmt mit dieser Krankengeschichte überein? (Wähle eine)

- 1 - hoher Blutdruck ist eine wahrscheinliche Ursache dieser Symptome
- 2 - Progesteronentzug ist eine wahrscheinliche Ursache dieser Symptome
- 3 - ein maligner Prozeß ist die wahrscheinliche Ursache dieser Symptome
- 4 - dies ist ein typisches Beispiel einer klimakterischen Blutung
- 5 - ein bösartiger Prozeß ist als Ursache der Symptome ausgeschlossen
- 6 - keine der genannten

Anmerkung: Die folgenden 4 Aufgaben gehören zusammen (173 - 176)

Die Patientin ist 47 Jahre alt. Sie gibt an, in den letzten 6 Wochen nahezu täglich aus der Scheide geblutet zu haben. Die letzte normale Periode hatte 8 Monate vor Beginn dieser jetzigen Blutungsperiode stattgefunden. Sie suchte ihren Arzt vor 3 Wochen auf, der einen Papanicolaou-Abstrich abnahm und eine Strichurettage zur Einschickung des Materials in ein pathologisches Institut vornahm.

- 1 - Röntgen- und Radiumbestrahlung
- 2 - radikale vaginale Uterusexstirpation (Schauta)
- 3 - erweiterte Konisation
- 4 - Amputation der Cervix
- 5 - Wertheimsche Radikaloperation

Später sagte er der Patientin, beide Tests seien negativ gewesen und ein Krebs daher ausgeschlossen. Er gab ihr "eine Spritze". Die Blutung sistierte jedoch nicht. Im übrigen hat die Patientin keinerlei Schmerzen und ist auch sonst in guter körperlicher Verfassung.

175 B

Welche der folgenden Feststellungen ist richtig in bezug auf die Untersuchung des Strich-Curettage-Materials? (Wähle eine)

- 1 - wenn ein Karzinom in der ersten Biopsie gefunden wird, muß noch mehr Gewebe untersucht werden, bevor die Behandlung eingeleitet werden darf
- 2 - ein negatives Ergebnis schließt einen Krebs aus
- 3 - ein negatives Ergebnis kann auch bei einem Karzinom vorkommen
- 4 - inaktives Endometrium zu finden, wäre bei der Anamnese dieser Patientin höchst unwahrscheinlich
- 5 - eine glandulär-zystische Hyperplasie in der Biopsie erweckt den Verdacht auf einen dysontogenetischen Ovarialtumor

Das unterschiedliche Abschneiden der Fragen aus der Kategorie "Stoffwechselstörungen" wird verständlich, wenn man sieht, daß von den sieben Fragen des Test A fünf unmittelbar oder mittelbar mit dem wichtigen Problem Diabetes zu tun haben, wie z. B. Fragen 131 A, während in Test B keine Fragen dieses Themas behandelt, dafür aber z. B. zwei theoretische Fragen wie die Items 137 und 138 bereits die Hälfte der Fragen ausmachen. Im übrigen finden die Ergebnisse dieses Subtests ihrer Entsprechung im relativen Abschneiden der Testformen A und B in der Querkategorie Endokrinologie-Metabolismus, welche die meisten der Fragen der Kategorie "Stoffwechselstörung" enthält und weiter oben besprochen wurde.

Fragenbeispiele aus den Subtests "Stoffwechselstörung"

Testform A M = 1.86
N = 7

131 A

Dringliche Sofortmaßnahmen bei diabetischem Koma sind:
(Wähle drei)

- 1 - intravenöse Traubenzuckerinfusion
- 2 - Salz- und Wasserersatz
- 3 - Insulininjektion
- 4 - Acidosebehandlung
- 5 - Infektionsprophylaxe

Testform B M = 2.50
N = 4

Anmerkung: Alternativen für die folgenden 2 Aufgaben

- 1 - eisenfrei
- 2 - eisenhaltig
- 3 - intrazellulär gelegen
- 4 - extrazellulär gelegen
- 5 - identisch mit Bilirubin
- 6 - mit Ferritin gekoppelt

137 B

Bei Zerfall von roten Blutkörperchen kann Hämosiderin entstehen. Welche der oben angeführten Eigenschaften kommen diesem Pigment zu? (Wähle drei)

138 B

Bei Zerfall von roten Blutkörperchen kann Hämatoidin entstehen. Welche der oben angeführten Eigenschaften kommen diesem Pigment zu? (Wähle drei)

6.2.4.4. Fragenkategorisierung nach Herkunftsfächern

Es wurde bereits eingangs auf die Problematik bei der Kategorisierung nach Fächern hingewiesen, da es sich hierbei um die Herkunft einer Frage handelt, die in manchen Fällen nicht voll mit dem Inhalt übereinstimmt. Darüber hinaus erwies sich die Einteilung in sechs Fachbereiche als zu grob bzw. die so gewonnenen Subtests als zu heterogen, um sinnvolle Aussagen aus möglichen Ergebnissen herzuleiten. Es soll deshalb nur kurz auf die beiden Verteilungen eingegangen werden, die sich als signifikant erwiesen und versucht werden, die Gründe für die in beiden Fällen negative Bewertungstendenz aufzuzeigen. Die schlechte Bewertung der Fragen aus der Pädiatrie in Test A beruht sicher auf einer ganzen Reihe von Aufgaben über seltene und praxisferne Problemstellungen aus der Genetik und dem Gebiet der angeborenen Fehlbildung (A 34 - 36). Außerdem fehlen in diesem Subtest fast völlig Fragen aus dem Gebiet der Infektionskrankheiten im Kindesalter, die zur relativ besseren Bewertung der Test B mit beitragen (25, 125).

Fragenbeispiele aus den Subtests "Herkunft Kinderheilkunde"

Testform A

M= 2.17

N= 30

Testform B

M= 2.03

N= 29

Anmerkung: Alternativen zu den folgenden drei Aufgaben.

- 1- XO; 45 Chromosomen
- 2- XY; 46 Chromosomen
- 3- XX; 46 Chromosomen
- 4- Trisomie -21; 47 Chromosomen
- 5- XXY; 47 Chromosomen
- 6- XY; 48 Chromosomen
- 7- Trisomie -17 und -18;
47 Chromosomen
- 8- 49 Chromosomen

34 A

Welche der oben genannten Chromosomenkonstellationen kann man bei einem 3 Jahre alten Mädchen mit Mongolismus finden? (Wähle eine)

36 A

Welche der oben angeführten Chromosomenkonstellationen findet sich bei einem 12 Jahre alten Knaben mit hypophysärem Minderwuchs? (Wähle eine)

25 B

Bei welcher der nachstehenden Krankheiten beginnt das Exanthem häufig in der Leistenbeuge? (Wähle eine)

- 1- Röteln
- 2- Scharlach
- 3- Varizellen
- 4- Masern
- 5- Exanthema subitum

125 B

Bei einem 2-jährigen Kind besteht seit 3 Tagen hohes Fieber und eine schmerzhafte Rötung und Schwellung im Bereich des rechten, unteren Oberschenkels und Kniegelenks. Alle anderen Gelenke sind unauffällig und frei beweglich. Wichtige diagnostische Maßnahmen sind: (Wähle zwei)

- 1- Röntgenaufnahme
- 2- Bestimmung der Leukozytenzahl
- 3- Bestimmung des Antistreptolysintiters
- 4- Blutkultur
- 5- Tuberkulinprobe
- 6- Knochenmarkpunktion

Als Ursache für die durchschnittlich schlechte Bewertung der Fragen aus der Pathologie in Test A sind eine Reihe theoretischer Fragen ohne ärztliche Anwendung (siehe Seite 50) anzusehen, wie z.B. Aufgaben 138 oder 132, 133. Demgegenüber zeigen ähnliche Aufgaben aus Test B engere Zusammenhänge mit ärztlicher Diagnostik oder Therapie (131/132/44).

Fragenbeispiele aus den Subtests "Herkunft Pathologie"

Testform A

M= 2.14
N= 22

Anmerkung: Alternativen für die folgenden 2 Aufgaben.

- 1- Milz, Leber
- 2- Niere, Nebenniere
- 3- Herzmuskel, Skelettmuskeln
- 4- Nerven
- 5- Lunge
- 6- Darm

132 A

Welche der genannten Organe bevorzugt die typische (die sekundäre) Amyloidose?
(Wähle drei)

133 A

Welche der genannten Organe bevorzugt die Paramyloidose (primäre Amyloidose)?
(Wähle drei)

138 A

Welche der Alternativen 1 bis 5 sind dem Abscheidungsthrombus zuzuordnen? (Wähle zwei)

- 1- seine Farbe ist rot
- 2- er liegt wandständig
- 3- er besteht hauptsächlich aus Erythrozyten
- 4- er entsteht postmortal
- 5- auf seiner Oberfläche finden sich die sogenannten Zahnförmigen Leisten

Testform B

M= 1.89
N= 28

44 B

Welches der nachfolgend aufgeführten Gewebe reagiert auf Röntgenstrahlung am empfindlichsten? (Wähle eine)

- 1- Muskulatur
- 2- Knochen
- 3- Epithel des Magen-Darmtraktes
- 4- Pankreas
- 5- Hypophyse

131 B

Welche der unten aufgeführten Vorgänge sind für die chronische Entzündung charakteristisch? (Wähle drei)

- 1- stärkere Proliferation des ortständigen Bindegewebes
- 2- arterioläre Hyperämie
- 3- Kapillarsprossung
- 4- Leukodiapedese
- 5- perivaskuläre Fremdzelleninfiltrate
- 6- seröse Exsudation

132 B

Welche der unten aufgeführten Hormone wirken entzündungshemmend? (Wähle zwei)

- 1- Thyreotropes Hormon
- 2- somatotropes Hormon
- 3- Desoxycorticosteron
- 4- Cortison
- 5- Hydrocortison

6.2.4.5 Fragenkategorisierung nach formalen Kriterien

Die Überprüfung von Zusammenhängen einiger Formalkategorien mit der Praktikerwertung erfolgte teilweise nach dem oben beschriebenen Mehrfeldertestverfahren, im übrigen in Form gewöhnlicher Korrelationsberechnungen sofern die untersuchten Merkmale normal verteilt waren. Auch im Falle der Formalkategorisierung wurde die 5%-Grenze der Irrtumswahrscheinlichkeit nicht erreicht oder unterschritten. Trotzdem sollen sich abzeichnende Tendenzen beschrieben werden.

Bezüglich der Fragenart erscheint die Anzahl der vom Durchschnittstyp abweichenden Fragen zu gering um aus den Ergebnissen der Mehrfeldertests annähernd zuverlässige Ergebnisse zu entnehmen. Wollte man dennoch eine Tendenz abzuleiten versuchen, so ergäbe sich, daß die Fragen vom Typ der "Weil-Frage" in Test A (z.B. 106) besser abschnitten als die Frage vom "Oberbegriff-Typ" in Test B wie z.B. Item 179.

Beispiele abweichender Fragentypen

"Weil-Frage"	M= 1.71 N= 7	"Oberbegriff-Frage"	M= 2.25 N= 4
--------------	-----------------	---------------------	-----------------

106 A

- A Jeder größere Ventrikelseptumdefekt sollte möglichst bald operativ verschlossen werden,
weil
- B am Defekt nicht selten eine Endokarditis entsteht

179 B

Liste I (Wähle einen Oberbegriff)

- 1- Plattenepithelkarzinom
- 2- Basaliom
- 3- Melanom

Liste II (Wähle eine Ausnahme)

- 4- Synonym: Ulcus rodens
- 5- Synonym: Basalzellenpapillom
- 6- keine Metastasen
- 7- große Strahlenempfindlichkeit
- 8- verhornendes Plattenepithel

Eine weitere Formalkategorie war die Fragengruppe, die einen "echten Fall" darstellen. Es sind dies jene Fragen, in denen versucht wird, die ärztliche Situation möglichst treffend zu simulieren, indem z.B. die individuelle Symptomatik eines Patienten geschildert wird. Auch hier zeigt die Untersuchung keine eindeutigen Befunde; jedenfalls läßt sich keine deutliche Bevorzugung dieses Fragentyps durch die Ärzte beobachten.

Als Grund hierfür muß die relative Häufung von Fragen, die klinische Probleme nicht aber Fragestellungen der Allgemeinmedizin behandeln, angesehen werden.

Fragenbeispiele aus den Subtests "echter Fall"

Testform A
M= 1.94
N= 34

Testform B
M= 1.76
N= 25

17 A

Ein 42-jähriger Mann mit einem reumathischen Herzleiden wird im Zustand der akuten Herzinsuffizienz ins Krankenhaus eingeliefert. Nach entsprechender Behandlung bessert sich sein Zustand dramatisch. Am Morgen des 3. Tages der stationären Behandlung hat er keine Zeichen des Herzversagens mehr. Am Abend des gleichen Tages bemerken Sie, daß der Patient Fieber hat.

Auf welche der folgenden Ursachen dürfte das Fieber mit der GERINGSTEN Wahrscheinlichkeit zurückzuführen sein? (Wähle eine)

- 1- Thromboembolie
- 2- bakterielle Endokarditis
- 3- aktive rheumatische Myokarditis
- 4- Herzinfarkt
- 5- Insuffizienz des Herzens

25 A

Ein 25 Jahre alter Mann klagt seit sechs Monaten über Schmerzen im Kreuz und an den unteren Extremitäten beim Aufstehen, aber auch während des Tages nach vorausgegangenen Ruheperioden. Bei der Untersuchung finden Sie:

Temperatur 36,8°C, Pulsfrequenz 84/min, Blutdruck 110/70 mmHg, eine leicht eingeschränkte Flexion der Lendenwirbelsäule, Verspannungen der anliegenden Muskulatur und Unterschied des Thoraxumfanges bei Expiration und Inspiration von 2 cm. Über dem Herzen ist ein leises diastolisches Geräusch mit punctum maximum über der Herzspitze zu hören. Lymphknoten, Milz und Leber sind

14 B

Ein junger Mann kommt 14 Tage nach einer Erkältung mit folgenden Symptomen zur Aufnahme: Hohes Fieber, blutgefärbtes Sputum, doppelseitige Lungeninfiltrate, leichte Hypertonie, Anämie, Hämaturie und Proteinurie. Er stirbt nach einigen Wochen im urämischen Koma. Die Sektion ergibt drei entscheidende Befunde:

- A keine Veränderung an den Arterien
- B diffuse intraalveoläre Blutungen mit Hämosiderosis der Lungen
- C eine schwere proliferative Glomerulonephritis

Die wahrscheinlichste Diagnose ist: (Wähle eine)

- 1- Kimmelstiel-Wilson-Syndrom (Glomerulosklerose)
- 2- Psittakose
- 3- Virus-Pneumonie mit Herdnephritis
- 4- Goodpasture-Syndrom (hämorrhagische interstitielle Pneumonie mit Nephritis)
- 5- subakute, diffuse, hämorrhagische Glomerulonephritis

118 B

Eine 45-jährige Frau wird wegen anhaltender Diarrhoen und einer bedrohlich zunehmenden Verschlechterung ihres Allgemeinzustandes in ein Krankenhaus eingewiesen. Zwei Jahre zuvor war sie am Magen operiert worden, und Sie erfahren vom behandelnden

nicht zu tasten. Die Blutsenkungsgeschwindigkeit beträgt 40 mm in der Stunde. Der Urinstatus ist normal. Die röntgenologische Untersuchung der unteren Rückenpartie zeigt geringe Veränderungen an beiden Ileosakralgelenken.

Die wahrscheinliche Diagnose ist:
(Wähle eine)

- 1- sogenannter Muskelrheumatismus
- 2- Arthritis urica
- 3- Morbus Bechterew (Spondylarthrititis ankylopoetica)
- 4- Nucleus pulposus Hernie
- 5- Spondylosis deformans

Arzt, daß wegen multipler und ganz ungewöhnlich lokalisierter Ulcera (im Oesophagus, Magen, Duodenum und Jejunum) eine subtotale Magenresektion vorgenommen worden war.

Sie legen eine Dauer-Magensonde und messen in 12 Stunden 3, 2 Liter Magensaft mit einem freien Säuregehalt von insgesamt 300 m Äq. Wie beurteilen Sie diesen Befund? (Wähle eine)

- 1- als normale HCl-Sekretion
- 2- als typisch für einen operierten Magen
- 3- als eine häufiger vorkommende Hypersekretion
- 4- als Zeichen einer seltenen, hormonell bedingten Hypersekretion

Zwischen der statistischen Schwierigkeit der Fragen - ausgedrückt in P = Prozent richtiger Antworten - und dem Urteil der Ärzte zeigte sich eine positive Korrelation. Diese Tendenz zeichnet sich für Test B deutlicher als für Test A und nur knapp unterhalb der 5%-Signifikanzgrenze ab. Für jeden Test existieren 2 P-Werte: Einmal auf der Basis von Experimenten mit englischsprechenden, zum anderen mit deutschsprechenden Studenten.

Tabelle 7: Korrelationen zwischen den Ärztebewertungen und der Fragenschwierigkeit "P" (= Prozent richtige Antwort)

	Test A	Test B
Allgemeinpraktiker / P englisch	+ .07	+ .12
Allgemeinpraktiker / P deutsch	+ .02	+ .14

Die Trennschärfekoeffizienten (Rpbi) der Fragen, die ein Maß für die Gültigkeit der Einzelfrage gemessen am ganzen Test darstellen, korrelierten negativ mit dem Urteil der Ärzte. Für die Korrelation der Rpbi-Werte aus englischen Experimenten mit dem Test B und dessen Bewertung durch die praktischen Ärzte ist dieser Sachverhalt ganz knapp (0.1447 gegenüber 0.1455) unterhalb der 5%-Signifikanzgrenze belegt.

Tabelle 8: Korrelationen zwischen den Ärztebewertungen und dem Trennschärfeindex (Rpbi)

	Test A	Test B
Allgemeinpraktiker/Rpbi englisch	-.10	-.14
Allgemeinpraktiker/Rpbi deutsch	-.05	-.07

7. Diskussion der Ergebnisse

7.1 Beteiligung der Ärzte

Die Beteiligung der Ärzte lag mit 66% hoch, wenn man bedenkt, daß die Bearbeitung des Tests mindestens drei Stunden, meist jedoch sicherlich noch mehr Zeit in Anspruch genommen haben dürfte. Damit scheint die Annahme bestätigt, daß von der ausgewählten Gutachtergruppe eine überdurchschnittliche Kooperationsbereitschaft zu erwarten sein würde. Man darf allerdings nicht vergessen, daß diese hohe Beteiligung sicher auch wesentlich auf den umfangreichen Kosten- und Zeitaufwand zurückzuführen ist, der zur Beibringung der Beurteilungen geleistet wurde. So gingen 40% der erhaltenen Bewertungen erst nach einer schriftlichen bzw. schriftlichen und telefonischen Erinnerung der Gutachter ein. Besonders die telefonische Kontaktaufnahme, die in 14 Fällen noch eine Bewertung brachte, welche sonst vermutlich nicht mehr zu erhalten gewesen wäre, erwies sich als wirkungsvolles Mittel, um Ärzte - doch noch - zur Mitarbeit zu gewinnen. Einige Telefongespräche bestätigten daneben die Befürchtung, daß die zeitliche Belastung der Ärzte einerseits und die zeitaufwendige Beurteilung andererseits in vielen Fällen den Praktikern eine Bearbeitung des Tests unmöglich machten.

7.2 Beurteilung des Tests und der Fragen

Bei der Betrachtung der Bewertung der Testfragen durch die Praktiker fällt zunächst die positive durchschnittliche Gesamtbewertung der Tests mit 66.3 für Test A und 64.8 für Test B auf. Dies mag erstaunen, wenn man bedenkt, daß die Fragen des Tests sämtlich von Universitätsprofessoren formuliert wurden, deren Tätigkeits- und Problembereich sich bekanntlich wesentlich von demjenigen der Allgemeinärzte unterscheidet. Es scheint also, daß Professoren - jedenfalls die Autoren der Tripod-Test-Fragen - offenbar trotz ihres jeweils recht speziellen Tätigkeitsfeldes recht gut in der Lage sind, Testfragen zu formulieren, die für die Erfordernisse der praktischen Medizin Gültigkeit besitzen. Dies schlägt sich ebenfalls in den hohen Korrelationen nieder, die sich ergeben, wenn man die Wichtigkeitseinschätzung der Testfragen durch Praktiker und deutsche oder englischsprachige Professoren als Gutachter vergleicht. Sie erreichen, wie berichtet, beim Vergleich der Praktiker-

urteile mit denjenigen deutscher Professoren 0.66 (Test A) und 0.75 (Test B). Betrachtet man dann die durchschnittliche Bewertung der einzelnen Fragen, so zeigt sich, daß eine deutliche und vor allem einhellige Trennung wichtiger von unwichtigen Fragen, wie sie erhofft worden war, im großen und ganzen nicht erfolgt ist. Dies läßt sich am anschaulichsten aus der Abbildung 5 auf Seite 33 entnehmen. Auf die U-Form der in diesem Koordinatensystem dargestellten Verteilung wurde bereits hingewiesen. Die Häufung von Fragen im Scheitelpunkt des "U" ist Ausdruck einer relativen Uneinigkeit der Gutachter bezüglich dieser Fragen, da dieser Scheitelpunkt zugleich auch der Ort hoher Abszissenwerte, also hoher Standardabweichungen ist. Der Rest der Fragen bildet die beiden Schenkel des "U", stellt also diejenigen Fragen dar, die mit relativer Einigkeit von den Praktikern für "außerordentlich wichtig" (oberer Schenkel) oder "gänzlich unwichtig" (unterer Schenkel) gehalten wurden. Der Raum zwischen den beiden Schenkeln indessen bleibt praktisch frei, was bedeutet, daß so gut wie keiner Frage von allen Gutachtern übereinstimmend eine mittlere Wichtigkeit zugeschrieben wurde. Anders ausgedrückt: Wenn Fragen eine mittlere Durchschnittswertung erhielten, so in aller Regel deshalb, weil ein Teil der Ärzte sie positiv, ein anderer negativ bewertete, nicht aber weil die Mehrzahl sie für mittelmäßig wichtig hielt.

Der Grund für diesen uns wichtig erscheinenden Befund mag darin liegen, daß es sich bei der Bewertung einer Multiple-Choice-Frage um einen relativ vielschichtigen Vorgang handelt. Es muß zunächst aus dem Fragen-Antwort-Komplex das Kernproblem einer Frage herausgelöst werden. Sodann muß sich der Beurteiler darüber klarwerden, welche spezielle Fragestellung im Zusammenhang mit diesem Problem die Frage behandelt, z.B. Diagnostik oder Therapie und weiterhin, ob es sich um praktisch relevante oder rein klinische Formen der Diagnostik oder Therapie handelt. Außerdem sollte sich im Optimalfall der Beurteiler über die formale Qualität einer Frage klar sein, wenn er sie bewertet.

Bereits diese nur grobe Skizzierung von Teilschritten, die zur Einschätzung einer Frage erforderlich sind, läßt erkennen, daß unterschiedliche Akzentsetzung verschiedene Bewertungen ein und derselben Frage durch einzelne Gutachter denkbar macht. So könnte sich beispielsweise

der eine Gutachter bei seiner Bewertung von der Wichtigkeit des Kernproblems leiten lassen, während ein anderer nur dann etwa eine positive Wertung vergibt, wenn auch der spezielle Aspekt des Problems, den die Frage anspricht, im Bereich praktisch-ärztlicher Tätigkeit liegt und ein dritter zusätzlich formale Kriterien in sein Urteil miteinfließen läßt. Die Richtigkeit der Annahme, daß solcherlei Unterschiede im Bewertungsverhalten für die große Zahl unterschiedlich gewerteter Fragen verantwortlich ist, läßt sich aus den vorhandenen Daten nicht beweisen, erscheint uns jedoch am wahrscheinlichsten. Der Grund läge in diesem Fall sowohl in Eigenschaften der Frage als auch in Eigenschaften der Gutachter.

7.3 Zuverlässigkeit der Gutachterurteile

Daneben bestünde die Möglichkeit, die Gründe für die Bewertungsunterschiede allein oder doch ganz vorwiegend in den Eigenschaften der Gutachter zu suchen. Das würde bedeuten, daß es sich bei den ausgewählten Gutachtern nicht um eine relativ homogene Gruppe bezüglich ihrer Auffassung über die Abgrenzung des Bereichs der allgemeinpraktischen Medizin gehandelt hätte, was eine Ausgangshypothese für die Untersuchung war. Hier zeigt die statistische Analyse, daß zumindest in quantitativer Hinsicht eine Polarisierung bei den durchschnittlichen Gesamtbewertungen des Tests durch die einzelnen Praktiker nicht festzustellen ist. Dies wurde mit der Verteilung der Abbildung 3 auf Seite 30 dargestellt. Hier hatte sich gezeigt, daß die Durchschnittswerte im wesentlichen im Bereich von sehr positiv mit niedriger Standardabweichung der Einzelurteile bis zu mittlerer Bewertung bei hoher Standardabweichung liegen. Das eine Beispiel insgesamt negativer Bewertung des Tests erwies sich als Einzelfall. Ebenso sprechen die doch recht hohen Interkorrelationen der Ärzte untereinander dagegen, daß prinzipielle Auffassungsunterschiede innerhalb der Sachverständigengruppe bestanden haben. In die gleiche Richtung weist außerdem das Ergebnis der Faktorenanalyse, die nur einen gemeinsamen Faktor aufwies.

7.4 Zusammenhänge zwischen kategoriemäßiger Zuordnung und Gutachterurteilen

7.4.1 Inhaltskategorien

Die Anwendung der Mehrfeldertests auf einzelne Fragenkategorien, diente dem Zweck, inhaltliche Hinweise dafür zu liefern, welche Themenkomplexe im einzelnen von den Ärzten für wichtig oder unwichtig gehalten werden. Wenn auch nach statistischen Kriterien dieser Teil der Untersuchung nur wenige signifikante Ergebnisse gebracht hat, so schienen uns die sich abzeichnenden Tendenzen doch interessant zu berichten und den Versuch einer logischen Erklärung wert. Es wurde bereits oben auf die relative Subjektivität sowohl z.T. der Auswahl der Fragenbeispiele als auch der Interpretationen hingewiesen. Die Erklärungsversuche einzelner Bewertungstendenzen wurden ebenfalls bereits im Ergebnisteil anhand von Fragenbeispielen gegeben. Das eindrucksvollste Ergebnis zeichnete sich hier in der Kategorie "keine ärztliche Anwendung" ab, die auch statistisch signifikant schlechter abschnitt als der Durchschnitt. Umgekehrt entsteht bei der Einzelbetrachtung der Fragenbeispiele der Eindruck, daß Fragen immer dann am positivsten abschnitten, wenn sie möglichst nahe einer ärztlichen Primärsituation entsprachen, wie z.B. die Diagnostik von Infektionskrankheiten, ihre Pharmakotherapie oder Prognose. Fragen dagegen die mit aufwendigeren und an klinische Bedingungen gebundenen Methoden verbunden sind, schnitten demgegenüber schlechter ab, ebenso wie die Fragen, die pathophysiologische, biochemische und erbbiologische Grundlagen betreffen, aber keine unmittelbare ärztliche Anwendung besitzen.

Die Tatsache, daß in sehr vielen Fällen für die gleiche Kategorie in den beiden Testformen A und B entweder schon in der Tendenz oder bei gleicher Tendenz quantitativ so unterschiedliche Ergebnisse auftraten, weist darauf hin, daß die gewählten Kategorisierungen häufig nicht geeignet waren, eine sinnvolle Charakterisierung zur Zusammenfassung homogener Fragengruppen abzugeben. Zum einen erwiesen sich die Kategorien als rein quantitativ noch zu weit gefaßt, zum anderen blieb in aller Regel der offenbar wesentliche qualitative Aspekt: "klinisch-theoretisch" oder "praktisch" durch die Kategorien unberück-

sichtigt. Hieran zeigt sich erneut, daß für die Beschreibung und auch Beurteilung von Prüfungsfragen eine Mehrzahl von Aspekten zu berücksichtigen ist (16), was die exakte quantitative Analyse einer inhaltlichen Gültigkeitsprüfung wie der vorliegenden wesentlich kompliziert, wenn nicht unmöglich macht und uns dazu bewog, der intuitiven Interpretation der intuitiven Gutachterurteile einen relativ breiten Raum zu widmen.

7.4.2 Formalkategorien

Die Überprüfung des Zusammenhangs von Fragentypen und der Bewertung durch die Gutachter erbrachte, wie berichtet, keine verwertbaren Ergebnisse. Auf die niedrigen Itemzahlen wurde auch bereits hingewiesen, was eine Interpretation von vornherein wenig sinnvoll erscheinen läßt. Dies gilt allerdings nicht für die Fragenkategorie "echter Fall", die in beiden Tests ausreichend vertreten war. Es ist zunächst verwunderlich, daß diese Frageform nicht positiver bewertet wurde, obgleich sie doch wohl am ehesten die Entscheidungssituation widerspiegelt, in der sich der Arzt bei seiner Praxis häufig befindet. Bei einem großen Teil der vorliegenden Fragen, handelt es sich jedoch um die Problemschilderung von Patienten in klinischer Behandlung, was offenbar den Ausschlag für die nur durchschnittliche Bewertung dieser Frage gab.

Die Zusammenhänge von Formalkategorien wie Schwierigkeit und Trennschärfe mit den Fragenbewertungen der ärztlichen Gutachter führt hin zu prinzipiellen testtheoretischen Problemen, die jedoch in diesem Rahmen nur gestreift werden können, weil die hier vorliegenden Daten keine ausreichende Basis für eine ausführlichere Erörterung darstellen können. Sie findet sich bei Kapuste 1974 (9).

Betrachten wir zunächst die positive Korrelation zwischen der Prozentzahl richtiger Antworten ("Schwierigkeit") und den von den Ärzten vergebenen Bewertungen der Fragen (vgl. Tab. 7, S. 70). Die Konsequenz aus diesem Befund, unformuliert in eine Empfehlung für die Konstruktion medizinischer Leistungsteste würde lauten, daß solche Tests möglichst viele "leichte" Fragen enthalten sollten. Dies steht im Widerspruch zu Forderungen aus der traditionellen Testtheorie,

welche im Zusammenhang mit bestimmten mathematischen Abhängigkeiten zwischen Frageschwierigkeit und -trennschärfe die Konstruktion eines Tests aus Fragen aller Schwierigkeitsgrade fordern. Diese Forderung ist jedoch nur da sinnvoll, wo es z.B. um die Messung normalverteilter psychologischer Merkmale geht.

Der ärztliche Beruf dagegen erfordert in vielen Fragen hundertprozentige Sicherheit. Es geht daher nicht an, von einer Normalverteilung ärztlicher Kompetenz bei Kandidaten einer berufsqualifizierenden ärztlichen Prüfung auszugehen und dementsprechend medizinische Leistungstests nach den Kriterien traditioneller Testtheorie zu konzipieren. Ein gültiger Test zur Messung ärztlicher Kompetenz im kognitiven Bereich wird vielmehr notwendig zu einem großen Teil aus "leichten" Fragen bestehen müssen.

"Leicht" bedeutet nun aber in diesem Zusammenhang nicht, daß es sich um logisch einfache Fragen handelt, die evtl. durch Raten oder einfache Kombination ohne medizinisches Fachwissen zu lösen sind, sondern daß in diesen Fragen Inhalte von allseits anerkannter Wichtigkeit behandelt werden, die infolgedessen entsprechend sicher gelehrt und gelernt werden.

Ganz im Einklang mit dieser Überlegung steht auch der Befund, daß die in der Testtheorie beschriebene Abhängigkeit von Schwierigkeit und Trennschärfe offenbar für medizinische Leistungstests nicht in der Form gilt, daß hohe Trennschärfen nur oder v.a. im Bereich mittlerer Schwierigkeit anzutreffen wären, während im Bereich der schwereren, wie auch der leichteren Fragen die Trennschärfe abnimmt (13). Ein solcher Zusammenhang müßte sich in einer Korrelationsberechnung der beiden Parameter in einer Nullkorrelation ausdrücken. Für den Tripod Test dagegen ergeben sich Korrelationskoeffizienten von .14 (Test B) und .20 (Test A). Das sind positive Korrelationen mit einer Irrtumswahrscheinlichkeit von knapp über 5% im ersten und weniger als 1% im zweiten Falle. Es finden sich also in diesem Leistungstest im Bereich leichter Fragen eher hohe Trennschärfeindices. Über einen analogen Befund berichteten auch Betke und Butenandt in ihrer Analyse der Ergebnisse von Semesterabschlußklausuren in Pädiatrie an der Universitätskinderklinik in München (2). Vor dem Hintergrund dieser Überlegungen und Befunde ist die positive Bewertung "leichter" Fragen durch praktische Ärzte also durchaus plausibel.

Interessant ist demgegenüber allerdings die Tatsache, daß Trennschärfeindices und Gutachterbewertungen nicht positiv miteinander korrelieren (vgl. Tab. 8, S. 71). Versucht man dies zu erklären, so muß man sich zunächst vor Augen halten, daß der Trennschärfeindex als Kriterium der Qualität einer Frage nur dann gelten darf, wenn ein Test mit inhaltlich gültigen Fragen bezogen auf die untersuchte Gruppe hinreichend homogen ist, d.h. wenn die Lehr- und Lernvoraussetzungen der Kandidaten für die betreffenden Inhalte weitgehend gleich waren.

Das wird jedoch umso weniger der Fall sein, je breitgestreuter die Themen eines Tests sind. Der Lernende wird dann infolge der allgemeinen Informationsüberflutung in den Bereichen, die über das unbedingt Notwendige hinausgehen, zu einem gewissen Grad selektiv lernen müssen. Die Art dieser Selektion wird dabei je nach individuellen Interessen unterschiedlich sein, ohne daß jedoch hiernach Studenten mit guten oder schlechten Leistungen unterschieden werden könnten.

Für Fragen aus diesem angenommenen Bereich des selektiv Gelernten ist der Test dann nicht mehr homogen, bezogen auf die untersuchte Gruppe und der Trennschärfeindex verliert seine Aussagekraft. Nun laufen die Fragen des Tripod Test über sechs große Fächer, sodaß man entsprechende Überlegungen sicher auch auf ihn anwenden kann. Ginge man also davon aus, daß eine ins Gewicht fallende Anzahl von höheren Trennschärfeindices, wie sie für Tripod Testfragen errechnet wurden, im erläuterten Sinne nicht aussagekräftig sind, so ließe sich die fehlende positive Korrelation zwischen Ärzteurteilen und Trennschärfeindices möglicherweise dadurch erklären.

Es läßt sich allerdings auch noch eine weitere Überlegung zur Erklärung der fehlenden bzw. in der Tendenz sogar eher (wenn auch nicht signifikant) negativen Korrelation der beiden Parameter anstellen. So gehören zum Inhalt der meisten Vorlesungen der Kurse neben der bloßen Vermittlung der praktisch relevanten Fakten immer wieder auch Vertiefungen beispielsweise in pathophysiologische oder ätiologische Feinheiten sowie Exkurse in Randthemen, die für die Diagnose oder Therapie einer Krankheit keine unmittelbare Anwendung besitzen mögen. Sie werden dennoch sowohl von den Unterrichtenden als auch von den Interessierten unter den Studenten geschätzt, weil sie eine Auflockerung der bloßen Aufzählung

von Fakten darstellen und das Verständnis der Zusammenhänge fördern.

Von den Allgemeinärzten dagegen, die sich vermutlich eher an den unmittelbaren Erfordernissen der täglichen Praxis orientierten, können Fragen solchen Inhalts für weniger wichtig gehalten und deshalb schlechter bewertet worden sein, obwohl ihre Trennschärfeindices hoch sein dürften, weil sie besonders von den interessierten und besser informierten Studenten richtig gelöst werden.

8. Zusammenfassung

Ein internationaler standardisierter medizinischer Leistungstest (Tripod Test) wurde in einer schriftlichen Befragung einer Gruppe von 75 berufstheoretisch und praktisch erfahrenen Allgemeinärzten (Mitglieder der Internationalen Gesellschaft für Allgemeinmedizin, IGAM) zur Bewertung seiner inhaltlichen Gültigkeit vorgelegt. Die Fragen sollten auf einer 11-Punkte Skala von 0 bis 100 eingestuft werden, wobei 0 als "gänzlich unwichtig" und 100 als "außerordentlich wichtig für die Allgemeinmedizin" definiert waren. 66% der angeschriebenen Ärzte bewerteten eine der beiden vorliegenden Parallelformen des Tests mit jeweils 180 Fragen. 24 Bewertungen der Testform A und 23 Bewertungen der Testform B sind Grundlage der hier berichteten Befunde.

In einer statistischen Analyse wurden Mittelwert, Standardabweichung und Häufigkeit der Wahl einzelner Skalenstufen pro Gutachter und pro Frage errechnet.

Die Korrelationen der Ärzteurteile untereinander, sowie auch mit bereits vorliegenden Bewertungen durch Hochschullehrer wurden berechnet. Ferner wurden die Fragen nach bestimmten inhaltlichen und formalen Kategorisierungen zu Subtests zusammengefaßt und mit Hilfe statistischer Verfahren untersucht, ob sich in einzelnen dieser Subtests Häufungen positiver oder negativer Bewertungen finden ließen. In einer qualitativen Analyse der einzelnen Subtests und ihrer Bewertungen durch die Allgemeinärzte wird versucht, Erklärungen für gefundene Bewertungstendenzen zu geben.

Die Ergebnisse der vorliegenden Arbeit zeigen, daß es grundsätzlich möglich und sinnvoll ist, Sachverständige der Allgemeinmedizin an der Gültigkeitsanalyse schriftlicher Examensfragen zu beteiligen. Als wesentliche technische Schwierigkeit erwies sich die zeitliche Überlastung der Ärzte,

die dazu führte, daß in zahlreichen Fällen eine Beurteilung nicht oder aber erst nach mehrmaliger Bitte erreicht werden konnte.

Die Gesamtbewertung des Tests durch die Ärzte fiel positiv aus. Unerwartet war die hohe Übereinstimmung der Ärzteurteile mit denjenigen von Hochschullehrern, welche den Test bereits früher ebenfalls beurteilt hatten.

Bei der Betrachtung der durchschnittlichen Bewertung der einzelnen Testfragen durch die praktischen Ärzte mußte für eine relativ große Anzahl von Fragen eine unterschiedliche Bewertung durch die einzelnen Praktiker festgestellt werden. Der Grund hierfür liegt nach der Auffassung des Autors vor allem in der Mehrschichtigkeit des Vorgangs der Beurteilung einer Testfrage, die bei unterschiedlicher Akzentuierung einzelner Aspekte zu divergenten Urteilen führen kann. Eine wichtige Rolle in diesem Zusammenhang scheint die Würdigung der Tatsache zu sein, ob eine Frage zu einem unstrittig wichtigen medizinischen Problem mehr die praktische oder mehr die dem klinischen Bereich zugehörigen Aspekte des Problems berücksichtigt. Hier scheinen einzelne Ärzte ihr Urteil allein von der Wichtigkeit des Problems, andere dagegen zusätzlich von Praxisnähe seiner Behandlung im Einzelfall abhängig gemacht zu haben.

Eine zweite mögliche Erklärung für die unterschiedliche Bewertung so vieler Fragen wären grundsätzliche Auffassungsunterschiede bei den Gutachtern über die Wichtigkeit einzelner Themenbereiche. Sie kann aufgrund der vorliegenden Daten nicht ausgeschlossen werden, erscheint uns jedoch die weniger wahrscheinliche Erklärung, wenn man die Interkorrelation der Ärzte untereinander, sowie die Übereinstimmung ihrer Bewertungen mit denjenigen von Hochschullehrern als Maß für die Zuverlässigkeit der von ihnen vergebenen Bewertungen betrachtet.

Der Versuch Gesetzmäßigkeiten im Urteilsverhalten der ärztlichen Gutachter in Bezug auf allgemeine Einteilungskategorien zu ermitteln, muß als im wesentlichen gescheitert angesehen werden, was die Signifikanzprüfung der errechneten Korrelation der Ärzteurteile mit den verschiedensten Kategorisierungen der Fragen betrifft. Der Grund hierfür liegt sicherlich in der Vielschichtigkeit des Bewertungsvorganges, der durch die angewandten Kategorisierungen nicht in angemessener Weise Rechnung getragen wurde, da offenbar das Gültigkeitskriterium nicht in der Kate-

gorie oder Form der Frage liegt, sondern Frage für Frage einzeln bestimmt werden muß. Damit soll nicht behauptet werden, daß andere Faktoren, wie verschiedene Handhabung unserer Bewertungsskalen durch die Gutachter, Skalierungsprobleme also, die in der Psychometrie systematisch behandelt werden, keine Rolle spielen können. Trotz allem konnte die qualitativ-intuitive Betrachtung der Tendenzen der Korrelationen an einzelnen Beispielen zahlreiche Erklärungen wahrscheinlich machen, die für die Konstruktion und Auswahl praxisnaher Prüfungsfragen Hinweise geben.

Abschließend kann festgestellt werden, daß bei der Einbeziehung Praktischer Ärzte in die Gültigkeitsanalyse medizinischer Testfragen im vorliegenden Experiment zwei Hauptprobleme sich herausstellten:

1. Die zeitliche Belastung praktisch tätiger Ärzte als Hindernis für die Kooperation
2. Die Mehrschichtigkeit des Bewertungsvorganges von Testfragen als Problem der Datenanalyse und -interpretation.

Das erste Problem ließ sich bei uns lösen, weil wir glücklicherweise Zugang zu einer ohnehin stark motivierten Gruppe von Gutachtern fanden und sehr um gute Kooperation bemüht waren. Wollte man eine solche Gruppe für eine kontinuierliche - und damit immer besser werdende - Mitarbeit an der Verbesserung ärztlicher Prüfungen gewinnen, müßte natürlich dafür gesorgt werden, daß nicht nur Idealismus sondern reale Bezahlung realer Arbeit Grundlage der Kooperation werden.

Das zweite Problem ließe sich sicher durch methodische Erarbeitung von Lösungen für alle zu berücksichtigenden Aspekte erzielen. Die Mehrschichtigkeit des Bewertungsvorgangs von Prüfungsfragen verlangt die multivariante Datenanalyse oder eine verlässliche Trennung der verschiedenen Beurteilungsebenen (16), oder aber das, wie uns scheint, berechnete Vertrauen in den Wert intuitiver Urteile von gut gewählten Gutachtergruppen. Dabei soll freilich hier nicht der Eindruck erweckt werden, als seien praktische Ärzte die einzig und allein kompetenten Gutachter für die Gültigkeit medizinischer Testfragen. Sicher wird es nötig und sinnvoll sein, auch andere Gruppen an dieser Gültigkeitsbewertung zu beteiligen. Wir denken dabei etwa an Ärzte in den poliklinischen Ab-

teilungen der Krankenhäuser, welche diejenigen Patienten weiterversorgen, die von den praktischen Ärzten als Notfälle eingewiesen werden. Unter diesen Fällen finden sich z.B. immer wieder typische Fehldiagnosen und damit zusammenhängende falsche Anbehandlungen. Die zur Vermeidung solcher Fehldiagnosen erforderlichen Kenntnisse wären natürlich überaus wichtige Inhalte für die medizinische Ausbildung und ärztlichen Prüfungen, werden jedoch vielleicht gerade von einem Teil der Allgemeinärzte nicht in der ihnen zukommenden Wichtigkeit gesehen.

Die Gültigkeitsbewertung ärztlicher Prüfungsfragen durch allgemeinärztlich tätige Gutachter ist jedoch in jedem Falle eine wesentliche Komponente der Konstruktion gültiger Meßinstrumente für die kognitive Qualität von Medizinstudenten und Ärzten.

Wir haben die aufgrund der hier berichteten Ärzteurteile besten 90 Fragen beider Testfassungen herausgezogen und eine Kurzfassung für Ärzte herausgebracht (12). Mit dieser Kurzfassung konnten aufschlußreiche Untersuchungen durchgeführt werden, über die an anderer Stelle berichtet wird (9).

LITERATURVERZEICHNIS

1. Approbationsordnung für Ärzte, 28.1.70, Deutscher Ärzteverlag GmbH, Köln 1971
2. Betke, K. und Butenandt, I.: Der Lehrerfolg im klinischen Medizin-studium: Schriftliche Prüfungen mit Mehrfach-Wahl-Fragen, MMW 115 (1973)
3. Dixon, W.J. (Edit.): Biomedical Computer Programs, University of California Press, Berkeley and Los Angeles 1967
4. Feest, J. und Kapuste, H.: Interviews in Ixburg, Urban und Schwarzenberg, München-Berlin-Wien 1970
5. Greenhill, S. and Singh, H.J.: Comparisons of the Functions of Medical Practitioners in Rural Areas with those in Urban Areas, JME Vol. 39, Nr. 10 October 1964
6. Guilford, J.P.: Fundamental Statistics in Psychology and Education, Mac Graw-Hill Book Company, Inc., New York - Toronto - London 1956
7. Gulliksen, H.: Theory of Mental Tests, Wiley, London 1967
8. Hubbard, J.P. and Clemens, W.V.: Multiple Choice Examinations in Medicine, Lea and Febiger, Philadelphia 1968
9. Kapuste, H.: Ein objektiver Test für die Kenntnisse des Arztes, black box 1, Nr. 3, Institut für Ausbildungsforschung 1974
10. Kapuste, H. und Noack, H.: Leistungstests in der Medizin, Instruk-tive Mitteilung Nr. 1, Institut für Ausbildungsforschung 1968
11. Kapuste, H., Noack, H. und Götz, A. (Hrsg.): Tripod Test, Ein inter-nationaler Test für Medizinstudenten, Institut für Ausbildungsfor-schung 1969, Form A und Form B, je 180 Fragen
12. Kapuste, H., Noack, H., Götz, A. und Schönhals, C. (Hrsg.): Tripod Test, Kurzfassung für Ärzte, Institut für Ausbildungsfor-schung 1971, Form A und Form B, je 90 Fragen
13. Lienert, G.A.: Testaufbau und Testanalyse, J. Beltz, Weinheim-Berlin - Basel 1969
14. Lindquist, E.F. (Edit.): Educational Measurement, American Council on Education, Washington D.C. 1966
15. Noack, H.: Über experimentelle schriftliche Prüfungen und ihre Be-deutung für die Reform der ärztlichen Ausbildung, Berliner Ärzte-blatt 11 (1967)

16. Schönhals, C. und Kapuste, H.: Medizinische Ausbildung und allgemeinärztliche Praxis, *Ärztliche Praxis* 47 (1971)
17. Schulz, U., Remschmidt, H. und Prinz, H.: Leistungstests im Medizinstudium, *Didacta Medica*, Beiheft (1970)
18. Täschner, K.L.: Untersuchung über die Anwendung von Multiple-Choice-Fragen zur Überprüfung des Wissens in Gynäkologie und Geburtshilfe, Inauguraldissertation, Frankfurt/M 1968
19. Wallace, D.: A Case for - and Against - Mailed Questionnaires. *Public Opinion Quarterly*, Spring 1954

Lebenslauf

Ich wurde am 22. 5. 1947 als 4. Kind des Werner Schönhals und seiner Frau Gerda, geb. Birke, geboren. Bedingt durch mehrmaligen Wohnungswechsel meiner Eltern, besuchte ich nacheinander folgende Schulen: 1953 - 54 Volksschule in Gehlenbeck, 1954 - 57 Willemer-schule in Frankfurt, 1957 - 60 Oberrealschule in Schweinfurt. Von 1960 an besuchte ich das Gymnasium in Wiesloch, wo ich im Oktober 1966 das Zeugnis der Reife erlangte.

Im WS 1966/67 begann ich das Medizinstudium in Heidelberg. Dort bestand ich im Oktober 1967 das Vorphysikum und im Februar 1969 das Physikum. Im SS 1969 begann ich den klinischen Abschnitt des Studiums in München, wo ich im Dezember 1972 das medizinische Staatsexamen bestand.

Im folgenden Jahr leistete ich meine Medizinalassistentenzeit mit je vier Monaten in der Kinderchirurgischen Klinik, in der Intensiv-abteilung des Kinderzentrums und in der I. Medizinischen Klinik der Universität München ab und erlangte am 31. 1. 1974 die Approbation als Arzt.

Mein Berufsziel liegt im Bereich der Kinderheilkunde und Kinder-
psychiatrie. Gegenwärtig befinde ich mich in der Weiterbildung in
der Kinderabteilung des Max-Planck-Instituts für Psychiatrie in
München.

Christian Schönhals

München im August 1974

